DOCUMENT RESUME

ED 238 687	SE 043 550

AUTHOR	Kosslyn, Stephen; And Others
TITLE	Understanding Charts and Graphs: A Project in Applied Cognitive Science.
INSTITUTION	Consulting Statisticians, Inc., Wellesley, MA.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Jan 83
CONTRACT	400-79-0066
NOTE	464p.; Document contains marginal legibility on some pages.
PUB TYPE	Books (010) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC19 Plus Postage.
DESCRIPTORS	*Charts; *Cognitive Processes; Communication Research; Experimental Psychology; *Graphs; Mathematics; Perception Tests; *Psychological Studies; Schemata (Cognition); Visual Learning; Visual Literacy; *Visual Perception

ABSTRACT

This book, describing the result of extended research on how charts/graphs convey information, develops a scheme for describing/analyzing information contained in graphs/charts. A psychological theory of knowledge of the reader and the mental events which occur in attempting to read a graphic display are the two focal points of the book. A comprehensive research program aimed at various levels of difficulty of charts/graphs intended for a wide range of uses and a review of most of the existing literature on charts and graphs (85 references listed) are provided. The literature review is used to develop the analytic scheme/theory and to justify the methods chosen by the research team. A psychological approach is taken because of recent advances in cognitive science which allow a modeling of visual interpretation as something more than a simple recording system. The authors indicated that this is the first attempt at a comprehensive application of applied cognitive science, proving the usefulness of the knowledge being built up, supporting cognitive science in the sense of building a technology on it, and allowing deeper insights into human products and how to make them better. One hundred and thirty figures and tables are attached. (JM)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

ED238687

UNDERSTANDING CHARTS AND GRAPHS: A PROJECT IN APPLIED COGNITIVE SCIENCE

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

NIE 400-79-0066

Year 03

January 5, 1983

Submitted to:

The National Institute of Education
Contracts and Grants Management Division
1200 19th Street, N.W., Room 808
Washington, D.C. 20208

ATTN: Martin Engel, Ph.D.

Submitted by:

Stephen Kosslyn
Department of Psychology
The Johns Hopkins University
34th & Charles Streets
Baltimore, MD 21218

Steven Pinker
Department of Psychology
Massachusetts Institute of Technology
Cambridge, MA 02115

William A. Simcox
Senior Engineer
Consulting Statisticians, Inc.
20 William Street
Wellesley Hills, MA 02181

Leon P. Parkin
Senior Engineer
Consulting Statisticians, Inc.
20 William Street
Wellesley Hills, MA 02181

SE043550

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
National Inst. of Education
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

UNDERSTANDING CHARTS AND GRAPHS:

A PROJECT IN APPLIED COGNITIVE SCIENCE

Stephen M. Kosslyn

Steven Pinker

with

Leon Parkin

William Simcox

CHAPTER 1: Introduction

It is often said that a picture is worth a thousand words. But this is not always true; many pictures are not even worth a dozen words. The worst offenders may be charts and graphs, pictures that are intended to convey information more effectively than could be done using words and numbers. But as anyone who has even glanced through the major national news magazines knows, charts and graphs often fall woefully short of this goal. This book is about the reasons why charts and graphs are all too often ineffective, uninterpretable or semi-interpretable pastiches at best serving to make a page visually interesting. The other side of this coin is, of course, the ways in which charts and graphs can be made to be effective, and much of this book focuses on this topic.

Consider Figures 1.1 and 1.2, which appeared in Fortune magazine and the American Scientist, respectively. What is wrong with Figure 1.1? Can you understand it clearly? Most people are quick to notice that the colors used to draw the functions are too similar, and most people are confused by the tapering shape. Now, what about Figure 1.2? This one is so awry that most people have difficulty simply discovering what the graph is about. But why? We could hazard guesses, but this clearly is not the best way to proceed. What we need is a systematic and well-motivated way of diagnosing the problems with specific displays. At the end of Chapter 5 we will return to Figures 1.1 and 1.2, and see what more can be said about their shortcomings.

INSERT FIGURES 1.1 AND 1.2 HERE

This book describes the results of an extended research program on how charts and graphs convey information. This program has two major foci which are played upon in the ambiguity of the title. By "understanding charts and graphs" we refer not only to the scientist's analysis of charts and graphs, but to the process whereby a reader comprehends them. Our first aim is to develop a scheme for describing and analyzing the information conveyed in charts or graphs. This scheme is designed to lay bare the particular problems inherent in a display, if any. This analytic scheme is focused on the chart or graph as an object in its own right, and its properties are described in terms of how the set of lines and marks on the page function as a complex set of symbols embodying information about objects or events in the world. The scheme is oriented around a set of principles that must be obeyed if a display is to be readily interpretable.

The second focus of the research program has been to develop a psychological theory of the knowledge in the head of the reader and of the mental events that occur when he or she attempts to read a graphic display. This theory, in part, provides the justification for the way we analyze charts and graphs, in that our analysis is supposed to tap the features of charts and graphs that make them relatively easy or difficult for a human reader to comprehend. Thus there is, in fact, an interplay between the two foci of the program, the analytic scheme and the psychological theory; the second provides the backdrop for the first.

The research program we describe here differs from all other work on charts and graphs in two important respects. First, it is comprehensive. We consider charts and graphs at multiple levels of description, from lines on a page to abstract mathematical symbol structures to concepts in a person's head, and we consider charts and graphs intended for a wide range of different uses. In addition, in the course of developing our analytic scheme and theory, we review most of the existing literature on charts and graphs and how people comprehend them. This review is not included solely for purposes of completeness, however; rather, we use findings in the literature to help us develop both the analytic scheme and the theory, partly by using the findings to justify the way we have chosen to procede as opposed to alternatives considered by others. Second, our system is firmly rooted in concepts developed in the study of perception and cognition. As noted above, even our analytic scheme is psychologically-oriented, and is intended to reveal the ways in which a given chart or graph is difficult for a person to interpret.

Why a psychological approach?

The psychology of the graph reader is a running theme throughout this book, and our emphasis on it is in fact the raison d'etre of the entire research project on which the book is based. To a psychologist, the worth of this approach is obvious, and we hope the fruits of our research as presented here will lead the nonpsychologists among our readers to a similar conclusion. But until now, few have explored the relations between the design of good graphs and the psychology of the people who must read them. A search through the psychological literature of the last century turns up only a handful of studies on graph reading, and the "how-to" guides for graph designers often do not seem to make even the slightest concession to the fact that the intended audience for graphs consists of humans rather than robots or Martians.

The reason for this failure of minds to meet, we feel, is fairly simple. In everyday life, it is natural to think of our eyes as simple recording systems, registering the world as it is. But even a moment's reflection (not to mention a century of laboratory research) can show that this analogy can be misleading. Consider the following examples. Cereal manufacturers can design boxes that look twice as big as their competitor's, but do not contain twice as many cornflakes. We readily notice a gain or loss of 5 pounds on a slim person, but are oblivious to a weight change on an obese person unless it is many times more extreme. Rows of reflectors on a dim highway, or formations of geese flying overhead, stand out perceptually as cohesive solid objects; animals or airplanes with blotchy camouflage are not seen as objects at all. Naturally, these biases built into our eyes and brains will not disappear when we look at graphs instead of birds, and obviously an effective graph designer will do best by being aware of these biases.

Incidentally, our knowledge of human vision has underscored not only its deficiencies in comparison with mechanical optical systems but also in many cases, its superiority, and in others, sheer differences in operation. Check-reading machines can record those odd-looking numerals at the bottom of checks, but unlike our eyes, these mechanical visual systems cannot make head nor tail of the names and dates printed at the top. People don't think shadows are parts of objects or that a tree lined up with a person is attached to him, but even the cleverest computer vision systems are prone to make such mistakes. And computers in general are indifferent as to whether a given set of numbers enter their data banks as a pattern of black marks on a page, a pattern of beeps over a telephone line, a pattern of holes punched on a card, or a pattern of movement of a joystick. But humans seem to prefer their numbers in graphic form, even though lists of numbers (or for that matter precise patterns of rising and falling tones) can contain identical information. These peculiar biases of ours, taken together, can shed light on the structure of our visual systems, a structure that makes us unique among optical information processing devices. The details of this structure, in turn, determine the ease or difficulty with which people with various sorts of training will extract various sorts of information from various sorts of graphs.

Applied Cognitive Science

This book is more than a psychological analysis of charts and graphs or a psychological theory of how they are comprehended, although it is both of these things. Our approach is of a very recent vintage, expanding beyond the traditional boundaries of the field. In fact, this book represents the first comprehensive program of its kind in the emerging discipline of "cognitive science." Cognitive science draws theory, methodology, and conceptual tools from linguistics, philosophy, and computer science, in addition to psychology. We have put to use many ideas and techniques from this broader discipline in developing our analytic scheme and processing theory. The core of our analytic scheme is drawn from basic distinctions in linguistics and some ideas developed in philosophy, and the backbone of our theory rests on concepts developed in computer science. Further, in addition to drawing on the psychological literature to buttress our empirical claims we rely heavily on methodologies developed in linguistics to test specific aspects of our ideas.

Thus, this book demonstrates how one can "cash in" on the abstract ideas that have been percolating in cognitive science. Demonstrations of the applicability of a body of knowledge are useful for a number of reasons. First, the value of obtaining the abstract knowledge is underlined if it can be put to use (especially if the uses are unexpected, clearly spinning off of the abstract knowledge per se and not a special effort to discover something useful). Nobody questions the value of studying physics, if only because of the bountiful harvest of technology from the pure research. Second, the mere fact that a technology can be built upon the fruits of pure research is another kind of evidence supporting the theories and general approach that guided the research. That is, one metric of evaluation of a theory is how well that theory not only explains old data and predicts new data, but how well it leads to the produc tion of useful phenomena or insights. And, of course, there is a third reason why finding applications of cognitive science in particular is a good idea; it promises to give us deeper insights into human products--such as charts and graphs--and how to make them better. That is, we hope to use our theories and general approach to research to tailor the things we use in everyday life such that they are maximally compatible with how we think and what we are. This book is one demonstration of how such an enterprise can proceed.

BACKGROUND

Even a casual perusal of the literature immediately convinces one that there is a real need for research on charts and graphs, and that there is a real need for a systematic approach to the topic. Research on charts and graphs is, in a word, scanty. *Psycholcgical Abstracts* lists about a dozen studies conducted in the last quarter-century, many published in esoteric sources. The available literature falls into three classes: "How to" books for graph makers, graphic tools for statisticians, and laboratory research which compares graphs to other media and investigates the comprehension of charts in graphs in general. We will consider this last category when relevant in the remainder of this book, but let us get a sense of the general run of the field by examining the other two classes now.

The largest category of treatments of charts and graphs is clearly the "How to" books (e.g., Brinton, 1919; Carrol, 1960; Haskell, 1920; Lutz, 1949; Rogers, 1961). These books typically divide graphs into different categories (e.g., line graphs, bar graphs, pie graphs, and pictograms), provide pointers on how to construct them (based primarily on the author's experience), and offer a few rules of thumb as to which graphs should be chosen to represent which types of information (e.g., "trends should be conveyed by line graphs, and proportions by pie graphs"). They also offer suggestions on improving the clarity of graphs (e.g., "if overlapping lines on a graph are cluttered together and hard to differentiate, expand the vertical scale, draw the lines in different colors, or place the lines on separate grids").

Although "How to" books may serve well as basic primers, their usefulness to the researcher is limited for the following reasons. First, although the conventional taxonomies of graphs reduce the variety of graphs to a more manageable number, they do not specify graphs in terms of the relevant psychological dimensions, a prerequisite to predicting how easily the graphs will be understood. Second, the rules of thumb on the visual clarity of a graph and its appropriateness for representing a given type of information are not based on empirical studies of graph comprehension. Rather, they are based on the intuitions of the author, which may be unrepresentative or contaminated by his or her professional prejudices. Furthermore, concensus among many authors over a set of rules of thumb may not be an adequate indicator of their soundness. These "How to" books follow each other's presentations closely, and they may simply be presenting an arbitrary, institutionalized conventional wisdom. Third, the rules of thumb describe comprehensibility in vague, global terms. The problem with this is that while readers may report that graphs constructed according to these dictates are easy to understand, the information they get from the graphs may be distorted in subtle but important ways--such as the reader seeing merely an increasing trend when the graph should specifically be depicting an exponentially increasing trend. Finally, the rules of thumb do not illuminate in any obvious way the cognitive processes involved in graph comprehension, which must be understood if psychologically-motivated principles of graph construction are ever to be developed.

Nevertheless, the "How to" books do have some uses. First, they provide large and varied samples of graphs giving us the opportunity to test the power of any analytic scheme (as will be discussed in detail in Chapter 5). Second, the rules of thumb may have heuristic value--if one rule seems particularly plausible, it can direct attention to one aspect of a graph and some operating principle of a cognitive component, suggesting an area of potential research. Finally, once a solid theory has been developed, a test of its adequacy can be made by returning to the rules of thumb, and noting how well the theory can explain the effectiveness (or lack thereof) of these rules.

A second source of insights on the use of graphic techniques comes from statisticians (e.g. Barnes, Pearson, and Reiss, 1955; Duntemann, 1967; Mullet, 1972; Tukey, 1971; Wainer, 1974; also see the "Teacher's Corner" feature of The American Statistician). The statisticians offer ways to graph data that make certain properties of the data salient. The psychological hypothesis underlying these graphing tools is that a statistical concept or parameter can be most easily grasped if it is displayed as a (preferably unidimensional) visual parameter like length or size. In addition, there are sometimes more specific hypotheses--for example, consider Tukey's (1971) suggestion that the human visual system is better adapted to judging the degree and type of scatter about a straight line than about a curve. This notion led Tukey (1971) to suggest that when one wishes to depict goodness-of-fit of an observed to a theoretical distribution, one should use a "hanging histogram" instead of a conventional one. In a hanging histogram, one end of each histobar is anchored at the line representing the theoretical distribution function and the other end "hangs" down toward the abscissa (see Figure 1.3). This allows one to assess all of the histobars relative to the same horizontal line--a task thought to be easier than gauging the scatter of the upper ends of the histobars about the curved line representing the theoretical distribution in a conventional histogram.

INSERT FIGURE 1.3 HERE

Unfortunately, these suggestions are not much more valuable for the present purposes than are those of the "How to" books. Their effectiveness is unknown, they are seldom tested empirically, and in one instance in which a suggestion was tested (Tukey's hanging histogram proposal, in fact), no advantage over conventional techniques was found (Wainer, 1974). In any case, since the techniques are designed for highly specific types of information, their relevance to the cognitive processes involved in comprehending graphs is unclear. The usefulness of the statistical tools, then, is similar to that of the "How to" books. They have a heuristic function, leading one to test their predictions (especially their specific predictions about visual classification processes), and to search for explanations for those predictions that are confirmed.

Thus, we are forced to rely on the empirical studies of graphic comprehension, which attempt to collect data supporting a given claim. Without such data we simply have no idea which notions should be taken seriously and which merely seemed like good ideas at the time. But first we need a way of making sense of the data, a way of structuring the issues and investigations that will allow us to draw out the practical impact of research findings. Thus, in the following chapter we develop a conceptual framework for characterizing charts and graphs. This framework is then used in the two following chapters, in which we review the empirical findings that bear on each of our operating principles. We will cast a somewhat critical eye on these findings, attempting to cull out those which are so methodologically flawed as to be of dubious value. In point of fact, most of the studies of charts and graphs in the literature are not much more useful than the "How to" books in the statisticians proposals; they too should best be regarded more as a heuristic source of suggestions than as the genuine foundation for a body of research. Many of these studies confound perception and memory; all simply tally errors rather than scaling perceived values of a graphed variable psychophysically; and most of the studies performed prior to the 1960's exhibit serious flaws in their design (e.g., failing to counterbalance order of presentation of conditions, using only a single set of data to be presented to the subject in each format, not informing the subject of what should be attended to in the graph, and providing ambiguous instructions). However, there are ample findings in the mainstream psychological literature that do bear directly on the perception and comprehension of charts and graphs, although they have not previously been regarded in this way. We will consider these findings and then implications in conjunction with new data we will provide along the way.

Using this book

This book can be read in two ways. Each chapter develops some ideas pertaining to graph communication in some depth, often reviewing a sizeable body of literature. We hope that the reader interested in actually studying charts and graphs or in further developing a scheme like ours will find these technical details important. For the reader interested in simply obtaining some practical guidelines to designing better charts and graphs, it might suffice simply to skim the chapters and focus on the conclusions. However, it is our hope that graph designers will become increasingly aware of the psychological makeup of the audience for their creations, and thus we would encourage graph designers to try to absorb the psychological rationales for the guidelines in addition to the guidelines themselves.

Overview of the book

This book has three distinct sections. The following chapter develops the analytic scheme, an integral part of which are a set of "operating principles" which must be adhered to if a chart or graph is to communicate effectively. The principles themselves are developed in detail in the second major section of the book. The initial principles stem from well-established facts about the human perceptual system such as those mentioned earlier in the introduction, and thus there is a substantial body of laboratory research that pertains to them. This literature is reviewed and the morals for the chart and graph maker are distilled. In addition, there are principles that do not emerge from the study of basic perceptual processes, but are revealed only when we examine how the eye and mind interpret charts or graphs per se. These principles are derived in part by using an inductive methodology that has proved highly successful in the study of human language. We gathered a large representative sample of charts and graphs, assessed how easy or difficult each one appeared to us, and treated these judgements as empirical data about graph readers (in this case, ourselves) in need of explanation. As a first step towards that explanation, we formulated the smallest set of principles we could find that concisely categorized the problems we experienced in interpreting the graphs in our sample. The data we marshall in support of these principles is akin to those used by linguists concerned with developing grammars. Such grammars are developed and tested by considering which strings of words form proper sentences and which do not, and why. Instead of sentences, however, we construct minimal-difference pairs of displays, with the difference between them reflecting a difference in the operation of a specific principle. If one display is clearly inferior to the other, we reason, and this inferiority is localized to that aspect affected by the principle in question, then this provides support for the psychological validity of the principle. In most cases the inferiority of one member of a pair is overwhelmingly obvious, and hence the reader's intuitions can be treated as a kind of data in their own right. But we have taken the precaution of collecting data from naive subjects to butress our claims about the effects of violating our operating principles. In the final chapter of this middle section, we summarize the details of a survey of a representative sample of charts and graphs from a wide variety of sources, giving the reader some sense of the most common sins committed by graphic artists. We also include here the short form of our analytic scheme, which can be used by anyone to evaluate a chart or graph. This scheme has been validated and assessed for reliability, as described in this chapter. The first chapter of the final section of the book consists of a description of a new psychological theory of graph comprehension; a theory of what we know when we know how to read a graph, how we use that knowledge when we read a particular graph, and how we attain that knowledge to begin with. The theory uses a large body of research on perception, cognition, and memory to integrate the conclusions of the previous chapters and to generate predictions for future research. Next we offer a set of guidelines--based on both the theory and the analytic scheme--for constructing charts or graphs. And in the final chapter we show how the present project can be generalized to the design and use of maps, diagrams and other sorts of visual displays.

CHAPTER 2: CHARACTERIZING EFFECTIVE GRAPHIC COMMUNICATION

Everyone has had the experience of opening a well-known national news magazine and puzzling over a chart or graph, trying to figure out what it is about and what it is supposed to be telling the reader. Often one can point to some aspect of the offending bit of graphics and say that those lines are too close together or that mislabeled axis is the root of the problem. But often one is not so sure exactly what is wrong and unable to tell the artist how to improve his or her work. In this chapter we develop a scheme for describing a chart or graph that has led to a systematic way of characterizing what is right, and wrong, about any given chart or graph. Because of the way the scheme was designed, it should be easily used to describe any unambiguous chart or graph in a straightforward way. When it cannot be easily applied, this is like a red flag waving, telling us that there is something wrong. We have developed--and tested, as will be described later--a set of principles that should be adhered to if a chart or graph is to be effective, and usually one of these principles (to be described shortly) has been violated when the scheme cannot be used easily.

Types of Visual Displays

There are numerous and varied ways in which people illustrate ideas or concepts. Cartoons, for example, can illustrate the artist's impressions by subtle variations of the thickness of a line (making a politician appear to have a heavy, caveman-like brow). Similarly, M.C. Escher's bizarre visions can force the viewer to see things in a new light. But these artistic uses of visual media are not the topic of this book. We are concerned with how quantitative information and relations among qualities are communicated graphically. These displays necessarily use symbols--marks that are interpreted in accordance with convention. There are common types of "symbolic" displays, which differ in terms of what information is communicated and how information is communicated.

Graphs are the most constrained form, with two scales always being required and values or sets of values being associated via a "paired with" relation that is always symmetrical.

Charts are less constrained because the entities being related are less constrained (they can be depictions, names, or numbers) and there is a wide variety of possible relations (practically anything). Nevertheless, charts have an internal structure, where entities must be visibly connected to other entities by lines that serve as links. These links can be labeled or unlabeled, directed or undirected, and need not simply pair entities.

Maps are unlike charts and graphs in that they are not entirely symbolic: a part of a map corresponds nonarbitrarily to a part of a territory that is pictured. The internal relations among parts of a map are determined by the internal relations of what is pictured. However, maps usually include a symbolic component (e.g., different colors representing different population), and labels are paired with locations by superimposing them.

Diagrams are schematic pictures of objects or entities. These can be picturable objects, such as parts of a machine, or abstract concepts, such as forces acting on the parts. A diagram is symbolic in that special symbols (e.g., cross-hatching to illustrate curvature) are used; a photograph is not symbolic because no "conventional" means of representation are exploited. Unlike charts and graphs, the parts of a diagram correspond to parts of some actual object or entity; and unlike maps, parts of diagrams do not represent locations of a territory.

Finally, tables are the least constrained and most general of the lot. A table can have words, numbers, or pictures. They can be arranged any way the designer wants (providing, of course, that the arrangement allows the reader to extract the necessary information--but we will get to this shortly). Instead of numbers representing the population of each state, the illustrator can present a map-like drawing where the size of the state represents the number. Note that appearances notwithstanding, such an illustration is not really a map: it uses the shapes of the states as labels; the actual spatial relations among the states is irrelevant for the purpose at hand. The states could be broken into four main regions, north, south, east and western regions, if it so suited the artist--or even listed in a column in alphabetic order. Tables, unlike charts and graphs, either have no internal organization or are organized globally. A set of balloons whose size corresponds to the amount a politician has talked a given day uses size as a numerical value, and the order of the balloons is irrelevant. In some cases, however, the relation to row and column headings is important; the immediate pairwise relations among entries always is irrelevant.

In this chapter we concentrate on a detailed treatment of graphs, and to some extent charts, for a straightforward reason: graphs are the most general form which at the same time very is constrained. That is, there are numerous different types of graphs--line, bar, surface, divided bar, pictograph--and yet the way they function to communicate information is well-structured. Although some display types, such as maps, are more constrained (the shapes must resemble those of the regions being represented), they are also less varied. We hoped that by understanding charts and graphs we would develop a system rich enough to encompass all of the types of displays. In this case we would for the most part simply "relax" various stictures for making a good chart or graph when considering making a good map, diagram, or table. Thus, our approach in this book will be first to understand the most structured and demanding cases, where graphs are used to communicate detailed information clearly and concisely. We then will turn to special cases, where only some subset of the complete information need be conveyed, and will consider variants on the standard graphic formats and varieties of other display types.

This chapter has three major parts. We begin by outlining the foundations of our analytic scheme. Following this, we present the analytic scheme itself, filling in more details about the basic ideas and how they were implemented (in particular, we introduce the "operating principles" here). Finally, we present two examples of how the scheme is actually used to analyze charts and graphs.

I. THE ANALYTIC SCHEME

A. TWO FOUNDATIONS

The analytic scheme has two deep taproots. The first is the literature on how humans process visual input, and the second is the so-called theory of symbols.

Visual information processing

A wide range of activities is interposed between that instant when you first fixate your gaze upon a visual display and that moment when you successfully extract some given information from it. The explosion of interest in cognitive psychology in recent years has given us a general framework for talking about these activities and has given us a rich body of literature concerning their operation. An effective visual display must not require use of mental operations people cannot perform, and must be easily dealt with using the operations we do not have at our disposal. Thus, it will behoove us to consider briefly now (but in more detail shortly) what is known about visual information processing, and then to consider how to use this information to diagnose bad displays and guide in the construction of good ones.

Insert Figure 2.1 About Here

Consider Figure 2.1, which is a very simple schematic of three main types of visual processing. The left most box represents "sensory information storage". The information present in an after image is in this kind of storage. It is very brief (for only a few tenths of a second) and contains virtually unlimited information during that time. The middle box is "short-term memory" (the word "memory" here is being used as in a computer's memory--a place where information is kept). The information stored here is usually accompanied by some conscious experience (such as of saying a word to oneself), and can be held in short-term memory by rehearsal (rote repetition). Information only stays in short term memory for a few seconds unless actively rehearsed, and only a small amount of information (about 4 groups of items) can be held in this store at the same time. Short term memory is important here because it is the locus when conscious re-organization and re-interpretation takes place, and its limitations severely affect what kinds of re-organization and re-interpretation can take place (as will be discussed shortly). Finally, the right-most box is "long-term memory". This memory stores a huge amount of information for an indefinite amount of time; your childhood memories, your telephone number, and the name of your favorite book are all stored here, as well as your knowledge of arithmetic and how various types of graphs (e.g., line vs bar) serve to communicate information.

In Figure 2.1 are schematized a number of properties of our visual information processing systems that affect reading charts and graphs (along with all other visual stimuli). Four of these properties pertain to how information is transferred from sensory-information storage to short-term memory (and hence into awareness). First, if the stimulus is too small or not contrasted enough with a background, you will simply fail to see it. The discriminability limits of the system must be respected if any further processing is going to happen.

Second, there are well-known systematic distortions in size and other properties of objects. For example, if you estimate the relative areas of two circles, you are very likely to underestimate the size difference. These distortions are reasonably-well understood and can be avoided or compensated for in a display (as will be discussed in the following chapter). Third, some aspects of a stimulus are given priority over others; we pay attention first to abrupt changes of any sort (e.g., heavier marks, brighter colors). Fourth, stimuli are organized into coherent groups and units by the time we become aware of them. Much of this organization is "automatic", not under voluntary control, and is determined by reasonably-well understood properties of stimuli (e.g., proximity of elements). The grouping imposed by these automatic processes must be respected if a chart or graph is to be seen the way a designer intends.

Given that information has been transferred from sensory-information storage, the next constraint we must consider is the capacity limit of short-term memory. If too much information must be held in mind at once, a person will be unable to perform a task. Thus, the complexity of a display will be a major factor in determining its comprehensibility. Once a display is in short-term memory, it is described. Tall bars, for example, are described as "large". A picture of a tall tree standing for a bar in a bar graph will be described both as a tree and as large. The description assigned here on the basis of the appearance of a display must correspond to one stored in long-term memory if it is to be interpreted correctly. And the way a display will be interpreted depends on which stored information is most closely associated with the description assigned to the display. If a line is described as "steep" it will be taken to represent a "sharp rise" in prices or whatever; if it is described as "shallow" it will be taken to represent a "slow rise" (even if it is the same information, just graphed on different-shaped axes!).

Finally, in long-term memory the major constraint is a person's knowledge. If a person does not know the meaning of a word, or of a pattern of lines forming the framework of display, he or she will have trouble associating the description of the display with the correct interpretation. In addition to general background knowledge, knowledge of the task at hand can have some important consequences: if the initial description of the display does not correspond to any stored information, knowledge of the task at hand can lead one to consciously re-organize the pattern, leading to new description and a new attempt to interpret the description against stored information. For example, if one sees a Star of David, one will organize that as two overlapping triangles. If asked whether there is a hexagon in the pattern, one will have to reorganize the pattern before seeing the hexagon in the middle.

The foregoing activities are relevant whenever one is trying to interpret what one sees. The details of these activities have yet to be specified (in chapter 6 we present one theory), but the basic kinds of operations seem clear enough. We certainly know enough about each operation and properties of the system to apply this knowledge to the designs of visual displays. The "Psychological Maxim" is straightforward: Do not design a display that overtaxes the human information processing system. The analytic scheme we have developed is in part a systematic way of discovering whether a given display has violated this maxim. And if so, our scheme is designed to reveal exactly how a display offends our processing abilities and exactly which abilities have been compromised.

Symbol systems

The second foundation of our analytic scheme is the theory of symbols. Some aspects of charts and graphs have nothing to do with the operation of the information processing system. They have to do with the very nature of how symbols operate. In the ideal case, a chart or graph will be absolutely unambiguous, with its intended interpretation transparent upon the first glance. One way to think about this sort of unambiguity is in terms of mappings between symbols and concepts. If the graphic display is treated as a complex symbol, then we want a unique mapping between it and one's interpretation of it. Goodman (1968) has characterized systems that have the property of unique bidirectional mapping between a symbol and concept as being "notational." These systems, such as musical notation, are much stronger than we need here. In them there is not only a single way of interpreting a given symbol, but there is only one symbol that can be used for any given information. Our requirement here is less stringent: given a symbol, there would be only one way to interpret it. Thus, for present purposes, there are two important uses of the basic ideas underlying notational systems.

First, we are concerned with the external mappings between the marks on a page and the interpretation of their meaning. It is important that the lines on the page be read as intended and have the intended effect on the reader. Second, there are internal mappings, which specify how marks in a chart or graph are paired with other marks; this is especially important when a key is used, indicating how labels should be paired with different lines.

In Goodman's scheme, the first distinction of importance for present purposes is between a "mark" (also called an "inscription"), a "character class," and "compliance class." A mark is a configuration of lines, such as "A". A character class defines which groups of marks will be classed as equivalent, such as "A" and "a". A compliance class is the referent, the semantic interpretation, of the character class, such as "first letter of the alphabet."

The distinction between a mark, character class and compliance class is useful in allowing us to contrast cases where marks do and do not map into a character class. If a physical mark maps directly into the compliance class, variations in the marks (such as weight of the lines used) are information conveying -- which need not be true if a mark merely *signals* a character class. The distinction between marks that map into a character class and ones that map directly into a compliance class is the distinction between a *sign*, which is arbitrarily related to the thing represented (e.g., "C" could have been used as another mark for the character "A"), and a *depiction*, in which marks are non-arbitrarily related to the represented information.

Kosslyn (1980) offers a set of formal criteria for distinguishing between marks that signify and marks that depict. Briefly, marks that depict have the following properties, none of which are *necessarily* shared by those that signify. First, every portion of the mark is a mark of a portion of the referent. The symbol "*" depicts a particular snowflake if every arm, e.g., "'", corresponds to a part of the snowflake itself. Second, the distance between all portions of the mark correspond to the distances between the corresponding portions of the object itself. Third, the lines used to inscribe a mark are not arbitrary. That is, given the foregoing two criteria, as soon as "'" and "-" are used in inscribing the mark used to represent "*", the size and position of the remaining lines of the mark representing the snowflake are determined. In contrast, any configuration of lines can be defined as an instance of a character class.

Goodman offers five distinct formal requirements for a "notational system." A notational system allows one to represent information precisely and unambiguously. English, then, is obviously not a notational system since ambiguous words or sentences are possible. Musical notation, however, meets the requirements of a formal notational system. Even though notational systems are stronger than we need for present purposes, it will behoove us first to consider Goodman's five requirements for a notational system here; following this, we will trim these requirements down to meet our present needs. Two of these requirements are syntactic, concerning only the properties of marks and characters, and the other three pertain to the semantic interpretation of the symbols.

The two syntactic properties are simply put. First, one should not be able to map a given mark into two different character classes. Goodman calls this property "syntactic disjointness." Second, one in principle should be able to decide into which character class a given mark falls. Goodman calls this property "syntactic finite differentation." In other words, the first requirement states that marks must be unambiguously interpretable in principle, and the second states that one should be able to tell one mark from another so that one can interpret a given mark. It is important for present purposes to note that the second requirement can be easily violated. Consider an example where lines of different lengths are used as marks and where any difference in length, *no matter how tiny*, affects the character class into which the mark is mapped. Now, in this case between any two marks an infinite number of others exist, and so too with any two characters. Given that no physical measuring instrument is infinitely precise, this kind of situation violates the requirement of "syntactic finite differentiation," since one cannot decide precisely which character class a given mark signifies. In this case, the representational system would be called "syntactically dense." An example of a syntactically differentiated system is a digital clock, where every reading on the clock (i.e., every mark) is distinctly identifiable and maps into one character class (and hence, the system is also syntactically disjoint). An example of a syntactically dense system is a dial clock with no tick marks. Now every position of a hand is a different mark, which signifies a different--although not uniquely decidable--character (time). This system is also syntactically disjoint because no mark maps into more than one character, although it is impossible to *identify* discrete marks.

The three semantic properties of a perfect "notational system" are concerned with the way in which one interprets the meaning of marks; in Goodman's terms, they are concerned with the way in which characters are mapped into compliance classes. The first two properties parallel the syntactic ones discussed above. First, two semantic categories (compliance classes) should not overlap so that they share members (as often happens in English). In other words, this "semantic disjointness" property proscribes ambiguous marks. Second, in a notational system one can identify the compliance class into which a given mark should be placed. That is, the system has "semantic finite differentiation." If one cannot decide which interpretation a mark should be given, the system is "semantically dense." So, for example, a digital clock is semantically differentiated because every reading has an identifiable meaning (and is semantically disjointed because each reading has only one interpretation). A tire pressure gauge, in contrast, is semantically dense because every reading on the continuous scale has meaning but one cannot assign a precise meaning to any given reading (because between every two readings are an infinite number of other ones, precluding precise assessment of an individual reading). However, if a tire pressure gauge is marked off in discrete intervals, and all readings within an interval have the same interpretation, now the system is semantically differentiated. Finally, the last semantic requirement is that all the marks of a given character class should have the same compliance class. Another way of putting this is that if marks can be mapped into a character class, the semantic interpretation is in terms of the character and not the marks directly.

For present purposes, we have found it useful to streamline Goodman's scheme considerably. We are interested in identifying cases in which there is a failure to have an unambiguous mapping between marks and meanings in a chart or graph. In all cases, when such a problem has been identified it can be ameliorated by changing the marks used in the chart or graph; even when a label is ambiguous, a new word or two can be substituted. Thus, we are not especially interested in pinpointing a lack of differentiation or disjointness at the level of syntax or semantics. Given that a relevant concept and distinction exist in the readership population, we can merely be concerned with being sure that the external mappings from mark to meaning are in fact unambiguous by ascertaining that the marks are differentiated and the interpretation is disjoint. For internal mappings, we will be concerned with part-for-part correspondences, which again requires differentation and disjointness of the relevant parts.

B. THE DESCRIPTIVE PROCEDURE

We have two broad classes of factors that must be considered when designing a chart or graph. A display must not overly tax our information processing abilities, and it must not be ambiguous or deficient in necessary information. We have designed a system for describing any given display that allows one to diagnose problems--either psychological or formal--with the display. The system can only be applied easily to a perfect display; when there is any problem in using the system, this is like an alarm sounding, serving to alert one to a problem. The particular problem is revealed by where the system breaks down, and the way in which it breaks down. The system has three components, the description proper, the diagnostics, and the evaluation.

1. Generating a Description

A description of a chart or graph is generated at three levels, and at each level the description is in terms of a set of components and relations among them, as described below.

a) Syntax, Semantics, and Pragmatics

We begin by describing charts and graphs with respect to three broad classes of properties. The *syntactic* properties are those of the lines themselves; here the lines are not interpreted in terms of what they represent but are treated as entities in their own right. In this case, configurations of lines are classified as falling into a set of "form classes," and the way these configurations are organized together is specified. In our analysis, these form classes correspond to the major "basic level" constituents of charts and graphs, as will be described in the following section. The semantic properties are the direct meanings of the configurations of lines, what they depict or signify. The semantic analysis is the literal reading of each of the components of a chart or graph and the literal meaning that arises from the relations among these components. Finally, the pragmatic properties characterize the ways in which meaningful symbols convey information above and beyond the direct semantic interpretation of the symbols. At the level of pragmatics in language, for example, the question "Can you open the door?" is not really comprehended as a question; rather, it is a request to open the door. The conveyed meaning in this case is quite different from the literal semantic interpretation; pragmatic overtones of visual displays hinge on the particular description assigned to the visual properties of the display (e.g., "steep" vs. "shallow" lines are seen as different, even if the same data is presented).

b) "Basic Level" Graphic Constituents

We distinguish among four "basic level" constituents of a chart or graph. Our notion of a "basic level" is directly analogous to how Rosch (1978) conceives of the notion of a "basic level" in categorization hierarchies. In categorization, the basic level is the one that is as general as possible while still having as similar members as possible. For example, "apple," and not "fruit" or "Delicious apple", is the basic level in the hierarchy of inclusiveness defined by those names, because that category captures the most exemplars that are still very similar; going down the hierarchy results in fewer exemplars in the category, say Delicious apples, whereas going up the hierarchy, to fruit, results in the exemplars not being very similar to each other. Similarly, our basic level graphic constituents seem to be the most general way of classifying the components of a chart or graph that still have a high degree of similarity among the different instances of the class. However, in our case the similarity is not in appearance, but in function, in the role a constituent plays in how information is represented in a display. The four constituents we use are called the framework, the background, the specifier, and the labels. These constituents are defined at the level of semantics, in terms of the information directly conveyed. Figure 2.2 serves to illustrate these basic level constituents for a typical chart and graph.

Insert Figure 2.2 Here

The framework. The framework "sets the stage" whereby the specifier material can specify the particular information being conveyed. The framework represents the kinds of entities being related (e.g., year and oil production), but does not specify the particular information about them conveyed by the display (e.g., the amount of oil per year). The framework often has two parts, defined partly at the level of syntax: The outer framework extends to the edges of the display and serves the role just described; the inner framework is nested within the outer one and often intersects elements of the specifier. The inner framework (often a grid or regular pattern of lines) usually functions simply to map points on the outer framework to points on the specifier. In most cases, the framework serves to organize the display into a meaningful whole at the level of syntax. In some charts, however, this is not true (e.g., see Figure 2.2), although the framework still functions semantically as described above.

The background. It is important to distinguish the framework from the background of a chart or graph. The background serves no essential role in communicating the particular information conveyed by a chart or graph. If any given background were removed, the chart or graph would still convey the same information at the level of semantics. Although any given background is not a necessary part of a chart or graph (often the background is blank), occasionally a patterned background, such as a photograph, can serve to reinforce the information in a chart or graph at the level of pragmatics (e.g., dead soldiers in a graph about the horrors of war); a patterned background can also interfere with one's ability to read a display, as will be discussed in detail shortly.

The specifier. The specifier conveys the particular information about the entities represented by the framework. The specifier usually serves to map elements of a framework (actually present or inferred by the reader) to other elements of the framework. In graphs, the specifier is often a line (serving to represent a function) or bars which pair values on the x and y axes specified by the framework. In charts, the specifier material is often directed arrows connecting two boxes or nodes.

The labels. The labels are alpha, numeric or depictive (i.e., pictures) and provide an interpretation for another line or part thereof (which is a component of either the framework or the specifier).

In addition to describing these constituents in terms of their syntactic, semantic, and pragmatic characteristics, we also describe the interrelations among the constituents. Much semantic information, for example, arises from the ways in which the components are physically juxtaposed. In addition to simply assigning lines to one of the three basic level graphic constituent classes, we also describe the constituents in terms of their subcomponents (for example, the framework of the graph illustrated in Figure 2.2 is composed of two lines that are organized such that one is horizontal and one is vertical and they meet at the lower left side of the horizontal line). The subcomponents are described in terms of simple "Gestalt wholes" (such as line segments) and the relations among them. At one time, we considered introducing a set of "primitive elements" and relations which would provide a fixed "alphabet of shapes" to be used in all our analyses. This proved very difficult to do, however, and proved to be totally unnecessary for our purposes. The wide variety of charts and graphs seems to preclude specification of a reasonably small set of discrete elements from which all charts and graphs can be constructed, but even if this were possible, the important variations seem to occur at what we have dubbed the "basic level" of organization into the graphic constituents noted above.

2. The Diagnostics

If a chart or graph is unambiguous and easily read, one should be able to assign an unambiguous description to it. Whenever one has difficulty in describing it, however, this is an indication that the display is flawed. At this point one tries to categorize the flaw using two classes of diagnostics.

a) Operating Principles

Many of the problems with a display can be linked to violations of principles that describe the operation of the human visual system. These violations can occur at each of three levels of description.

Syntactic Principles. These principles describe constraints on how lines may be interpreted and organized. A syntactic problem is not tied to the lines having a specific meaning, but hinges on problems with extracting any meaning from lines. If these principles are violated, one either cannot read a chart or graph (without, perhaps, the aid of a magnifying glass and ruler), will systematically distort information when reading it, will tend to have difficulty organizing it correctly, or will find it difficult to hold the number of relevant lines in mind at once. These principles summarize what we know about how information is transfered into, and retained in, short-term memory.

Semantic Principles. These principles describe constraints on the ways patterns of marks are interpreted. A semantic principle is tied to how a specific meaning can be extracted from a configuration of marks. These principles were derived primarily through a review of the literature on how people will spontaneously describe a visual display, and on the kinds of concepts people must have to understand charts and graphs. If these principles are violated, people may become confused in interpreting the meaning of a chart or graph.

Pragmatic Principles. The pragmatic principles describe the ways people in our culture customarily import more meaning than is actually conveyed on the page or the ways in which context interacts with display comprehension. These principles were derived through an analysis of a set of charts and graphs, as will be described later in the book.

b) Formal Principles.

These principles are special-purpose formulations of those underlying Goodman's concept of a notational system. They describe aspects of charts and graphs that must be respected if the chart or graph is to be unambiguous.

3. The Evaluation

The final basic idea of our analytic scheme is that charts and graphs are created with a specific purpose in mind; they are intended to allow a reader to answer certain questions and not others. Thus, although an operating principle may be violated the chart or graph may not be impaired -- it may still be able to serve its purpose adequately. For example, the graph illustrated in Figure 2.2 violated what we will call the Principle of External Mapping (a formal principle) because the points on the function do not correspond unambiguously to points on the axes. But this is not an impairment in the graph, given its purpose. In fact, when graphs are used as idealizations to present a general principle, the additional information necessary to totally disambiguate the display may distract from the purpose (see our principles of processing priorities and limitations, to be discussed shortly). Thus, although our scheme faithfully exposes every little detail that violates an operating principle, not all of these violations may be important. Whether a violation of a principle renders a display ineffective depends on the purpose to which the chart or graph will be put. The scheme errs on the side of being too conservative, leaving it up to the human user to discount particular violations as he or she sees fit. This was the only real option, given that all other alternatives run the risk of not exposing potential problems with the chart or graph. Later in the book we will present a detailed theory of how people actually comprehend visual displays which will then guide us in applying the principles themselves. Before developing and using such a theory of information-processing, however, it will behoove us to explore the usefulness of the general approach being taken here.

II. USING THE ANALYTIC SCHEME

Our scheme produces a description of any given chart or graph at three distinct levels of analysis, the syntactic, semantic, and the pragmatic. The description revolves around characterizing the basic level constituents noted above, namely the framework, background (if present), specifier and labels, as well as the relations among them. In the course of discussing how the scheme assigns descriptions to charts and graphs, we will introduce the operating principles. These principles will be described only briefly here; in later chapters we will flesh out the details of each principle. This exercise will provide detailed guidelines for evaluating displays and also will provide requirements on a theory of how people process visual displays, which will be presented subsequently. Following this, the theory will then be used in conjunction with the descriptive scheme to provide guidelines for a design of a chart or graph intended to make a particular point.

The following is a description of how the scheme is applied to a single chart or graph. At the outset, however, we ask whether the chart or graph is in fact composed of a number of subcharts and graphs. That is, we ask whether there is more than one chart or graph present and whether there are systematic relations among the information in each. If so, the scheme is applied to each one separately and then to the set of charts and graphs together. An example of an analysis of a complex multipanneled graphic display will be presented in the final section of this chapter.

In each of the levels of analysis, we ask a number of questions that should be easily answered if the graph is unambiguous. If we have trouble arriving at a straightforward answer to any of these questions, this alerts us that one or more of our operating principles has been violated. We then simply consider each principle relevant to that level of analysis, checking to see if it has been violated. Thus, because the system is set up to reveal violations of these principles, it will behoove us to begin each section with a brief overview of the relevant principles themselves. Following this, we will consider the actual mechanics of generating a description of a chart or graph.

THE SYNTACTIC ANALYSIS

Operating Principles

We posit three broad classes of operating principles at the syntactic level that cannot be violated if a chart or graph is to be effective. Each of these classes contains a number of specific principles which themselves have specific aspects, as will be developed in detail in the following two chapters.

A. Principles pertinent to seeing the lines

The visual system imposes numerous constraints on how marks can be used to convey information in charts and graphs. The first set of principles bear on how lines, colors, and regions are accurately discriminated--which is a necessary prerequisite for further processing. We posit two principles that bear on the process of discriminating marks:

1. The principles of adequate discriminability

Variations in marks must be great enough to be easily noticed. These principles have two aspects:

a) Relative discriminability: Two or more marks must differ by a minimal proportion to be discriminated. The laws governing the size of this difference have been worked out for many types of marks and these laws comprise this principle, as is described in the following chapter.

b) Absolute discriminability: A minimal magnitude of a mark is necessary for it to be detected. This "absolute threshold" has been computed for many types of marks, as is described in the following chapter.

2. The principle of perceptual distortion

The visual system often systematically distorts the magnitude of marks along various dimensions (such as area and intensity). This distortion is described by the value of an exponent in a formula developed by S.S. Stevens and his co-workers, as is discussed in the following chapter. Marks can be intentionally altered to compensate for the distorting properties of the visual system (which, for example, make increases in area seem smaller than they are).

B. Principles pertinent to organizing marks into units

Marks are rarely seen as isolated dots on a page. Rather, individual marks usually are organized into perceptual units, such as occurs when a series of marks like "----------" are seen as forming a single line, not as a series of isolated dashes. A set of principles describes the main factors that determine which marks will be grouped together into a single perceptual unit. If these principles operate to group together elements of a display inappropriately, the display must be changed.

1. The Gestalt principles of organization

The Gestalt psychologists, who had their heyday during the 1930's, discovered almost 120 distinct laws that dictated how forms were organized. The more important laws (for present purposes) can be summarized by four general principles:

a) Good continuity: Marks that suggest a continuous line will tend to be grouped together. So, "----------" is seen as comprising a single unit, not 10 separate ones.

b) Proximity: Marks near each other will tend to be grouped together. So, "xxx xxx" is seen as two units whereas "xx xx xx" is seen as three.

c) Similarity: Similar marks will tend to be grouped together. So, "XXX@@@" is seen as two units.

d) Good form: Regular enclosed shapes will be seen as single units. So, "[]" is seen as a unit whereas "[-" is not.

2. Principles of dimensional structure

Marks vary along a number of dimensions, such as hue, size, height, and so on. Some of these dimensions cannot be processed independently of others. For example, it is impossible to see the hue of a mark (i.e., its shade of color, roughly) without seeing its saturation (i.e., the richness of the color, roughly). Thus, some dimensions are organized into single units whereas others (such as hue and height) are not. The dimensions that are "stuck together" in processing are called integral dimensions and the ones that are processed independently are called separable dimensions.

C. Principles of processing priorities and limitations

The visual processing system has quantitative and qualitative limitations. Partly because only a limited amount of information can be held in mind at once, some marks will be given priority over others. The information conveyed by these marks should be central to the display's message. Further, some kinds of comparisons are difficult for the visual system to perform, and hence a display should not require use of them. These facts are the basis for two kinds of principles:

1. Principles of processing priorities

Some colors, weights of line, and sizes are noticed before others. For the most part we do not have formal rules for determining which these are, but instead rely primarily on a general principle: the visual system is "a difference detector". Any sharp contrast will draw attention. In addition, some stimulus properties have been determined empiriclly to be "salient" (e.g., all other things being equal, a yellowish-orange is noticed before a deep blue). Physical dimensions of marks should be used to emphasize the message, not to distract from it (e.g., by making the background too prominent).

2. Principles of processing limitations

These principles fall into two categories:

a) Finite capacity: Only about 7 units can be seen at a single glance, and only about 4 can be held in mind at once. Graphic displays should not contain any unit (e.g., group of lines) which itself contains more than 4-7 subunits (e.g., lines).

b) Unit binding: It is more difficult to see and compare parts of perceptual units than it is to see and compare entire units. For example, "-" is more difficult to compare to the lower left leg of "x" (not a natural unit) than to "/" (a natural unit). Graphic displays should not require readers to decompose natural units in order to extract specific information, as occurs if single points along a time line must be interpreted.

Applying the Analytic Scheme

The main point of describing a chart or graph using our scheme is to reveal violations of the operating principles that impair the effectiveness of the chart or graph. In order to do this, however, one must generate a description of exactly what is out there, exactly how a chart or graph is composed.

Thus, our scheme requires one to engage in two distinct activities. First, one actually describes the chart or graph. This is especially the case at the syntactic level. Second, one asks questions about the description, checking to ensure that the description is unambiguous and transparent. If not, one or more of the principles has been violated. The level of detail of the description proper is motivated by the kind of information one will need later on to assign a semantic interpretation, and then the pragmatics, of the chart or graph--again with an eye toward discovering violations of the respective types of operating principles.

We begin by isolating the four basic-level constituents and then asking the following questions about them:

The Background

We first ask whether there is a background and, if so, we describe it. A background extends beyond the framework and does not actually help to convey the information in the display; removing the background would not impair how the chart or graph functions to represent information. Some backgrounds, however, can consist of patterns that make it difficult to detect the pictoral material or other lines (and hence, violate the principle of adequate discriminability). Other potential problems with background information will be discussed later.

The Framework

Next we examine the outer framework. We define the outer framework as the set of lines that serve to define the general entities that are addressed in the display.

> What are the elements? Are they lines? If so, of what shape, weight, and color? Are they clearly discernable?
>
> If lines function as axes, are they dense or differentiated?

We note whether the framework and its individual component parts are easily identified.

> How are the elements organized? Are the relations among the different parts clear? Does the organization violate any of the natural organizational principles? How many elements must be held in mind at once in order to organize them into the entire framework?

Occasionally, a chart or graph will also include an inner framework, such as lines that cross-hatch the interior of a chart. If there is an inner framework, the same questions noted above are asked of it.

Next, we consider the organization of the two frameworks, if both types are present. In particular, we ask how the similarity, proximity, and continuity of framework elements imply organization. Following this, we ask a number of general questions about the entire framework:

> Does the framework represent 2D or 3D space? Are quantities distorted because of an ambiguity here? Is color employed in the framework; if so, what is emphasized? If line weights are varied, what is emphasized? (This will be important later in our pragmatic analysis). What is the aspect of the axis? (That is, which axis is longer; this also will be important in the pragmatic analysis).

The specifier

We begin by isolating the class of visual continua used to represent information. We then describe how shape, size variations, color and texture are used. In a typical chart or graph, such as that illustrated in Figure 2.2, the line serving as a function cannot properly be described as being syntactically or semantically differentiated. Thus, this would seem to preclude the graph being unambiguous; recall that one of the properties of notation systems is differentiation, ensuring unique mapping from mark to compliance class. However, one must take two factors into account here: First, what is the intended use of the chart or graph? For many purposes only a rough approximation is desired, especially when graphs are idealizations (such as Figure 2.2) intended to illustrate some general point. Second, even when precise information is being conveyed, one is in fact working with "psychological units" of limited precision: our perceptual apparati simply cannot make discriminations beyond a certain limit. Thus, if the smallest discriminable segments of a line used as a function map unambiguously onto the smallest discriminable segments of the axes, the chart or graph can function notationally. Thus, we go on to ask:

What are the elements used to compose the specifier? Is it clear whether parts are overlapping or contiguous? Are there too many elements to keep in mind at once? Are variations used to convey information clearly distinguishable?

How are the elements organized? Is the organization clear? If the specifier does not clearly imply a 2D shape, does an ambiguity in the dimensionality preclude easy reading of the information?

Labels

We first consider three kinds of labels independently, and then turn to an analysis of the relations among the labels. We pay special attention to the title, asking first if there is one. If so:

Is the title clearly discriminable as a title?

What is the relation of the title to other elements of the chart or graph? Does it naturally tend to be organized such that it incorrectly appears to label only a local part of the chart or graph?

Next, we consider whether there is a remote legend or key. If so, we ask:

Is the information clearly readable?

Does the legend clearly separate itself from other elements of the chart or graph?

Is there too much material to be easily held in memory?

Depending on the type of labels used in the title and legend or key, the following questions are then asked about them (as well as about all other labels of each type).

Alphabetic labels: Are alphabetic labels present? If so:

Are they clearly readable?

How many are present?

What size of typeface is used for each of the labels? Note if different sizes are used for different labels (this may be important at the pragmatic level).

How do labels group together? Is the natural grouping congruent with the intended interpretation?

Numeric labels: Are digits used as labels? If so, ask of them the same questions asked of the alpha labels.

Depictive labels: Are pictures used as labels? If so:

Are they clearly identifiable?

How many are present?

Are they all the same size? (note differences)

How do these labels group together? Is the natural grouping congruent with the intended interpretation?

If color variation is an important component in the labels, are variations clearly discriminable?

Organization among the different types of labels

How are the labels organized? Do any natural organization principles result in an incorrect organization of the labels? (for example, does dissimilar typeface cause one to separate labels that should be grouped together? Does proximity of labels cause one to group them improperly? Are labels ordered in such a way that you group them improperly?)

Organization among framework and specifier

What is the relationship between the framework and specifier? Is the specifier completely contained within the framework?

Are lines of the inner framework confusible with the specifier?

Do natural organizational principles cause one to group the framework and specifier incorrectly?

If the dimensionality of the space is not 2D, is it consistent between the framework and specifier?

Organization among framework and labels

The organization between the framework and each type of label is considered separately, with the following information being provided (as appropriate):

How are the labels associated with the framework and parts thereof? Are value markings indicated along the framework? If so, do the labels clearly indicate the correct values corresponding to the associated portion of the framework?

Do any natural organization principles result in an incorrect organization of the framework and labels?

Organization among labels and specifier

How are the labels and specifier associated? Is all specifier labeled? If the label is remote, in a key, is the mapping from elements in the key to the specifier clear?

Do any natural organization principles result in an incorrect organization of the labels and specifier?

Organization among labels, framework, and specifier

Is too much material present to apprehend all at once?

Is too much material in too small an area?

Do natural organizational principles impair discerning the incorrect relations among the constituents?

THE SEMANTIC ANALYSIS

In considering the semantic content of a chart or graph, let us begin by briefly outlining the four operating principles we have posited, and then turn to our scheme for describing the semantic information in charts and graphs. It is at this level that the differences between some classes of charts and graphs as such become important, requiring us to develop two different sorts of semantic interpretations, one based on qualitative relations and the other based on quantitative relations.

Operating Principles

We have posited two classes of semantic principles, both of which are supported not only by ample findings in the psychological literature, but by new data we have collected (examples of the problems that arise when the principles are violated will be illustrated in chapter 4). These principles are concerned with the kind of description that will be assigned to a display and how it will be interpreted.

A. Principles of surface compatibility

The mark used to symbolize or depict an object or class must be appropriate for that role. Some marks inherently look like something other than what they are intended to represent, which impairs correct interpretation of them. This principle has three aspects:

1. Principle of representativeness

All marks have a preferred interpretation. The intended meaning of a mark should not conflict with the spontaneous interpretation of it. Thus, labels should name words that are indicative of the class (including the correct connotations) and pictures should depict appropriate objects (a picture of a penguin-like bird should not be used to label birds in general). In short, a label or picture should be of a representative or typical example of a class or of the class directly.

2. Principle of congruence

This principle has four aspects:

a) Description conflict: The description of the lines themselves should be compatible with their meanings. For example, for words printed in different colored inks, people have trouble reporting the color of the ink if the words themselves name different colors (e.g., the word "red" is printed in blue ink; this is known as the "Stroop effect"). Thus, larger symbols (described as larger) should represent larger qunatities, faster rising lines should represent sharper increases, larger typeface should correspond to larger objects, and so on.

b) Aligning Dimensions: The "more" and "less" poles of a dimension used in a graph should correspond to the "more" and "less" poles of the variable it represents, respectively. Thus light patches should represent smaller quantities, and dark patches greater quantities, rather than vice versa; similarly, marks that are high, tall, wide, long, saturated, filled, dense, or sharp should represent larger rather than smaller quantities. If in doubt, say the words for the two poles in each order; the pole that is first in the better sounding order is the "more" pole (e.g., long and short sounds better than short and long, thus long is the "more" pole.

c) Markedness: Some words name not only a pole of a dimension but the dimension itself. We say "how high is that?" without implying necessarily that it is high; but if we say "how low is that?" we imply it is low. The term that implies a specific value is called the marked term, and should not be used to label the dimension itself--if it is, it will mislead the reader. Similarly, one should use the unmarked member of a pair of comparative terms: "larger" is better than "smaller", and so on.

d) Principles of cultural convention: The conventions of a reader's culture should be obeyed when drawing an effective graphic display. So, for example, the color red should not be used to represent "safe" areas, and green should not be used to signify "danger." Similarly, time should increase going left to right or bottom to top.

B. Principles of schema availability

In order for a chart or graph to be comprehensible, a reader must have the requisite concepts. That is, a "compliance class" is in fact something in a reader's head. The reader must know both the individual concepts and the general idea of how a particular graphic design conveys information.

1. Principle of concept availability

A chart or graph should not make use of concepts that are not likely to be possessed by the intended readership.

2. Principle of graph schema availability

Information should not be presented in a graph type that is unfamiliar to a given readership or that taxes the information-processing abilities of the readership population.

The Formal Principles

In the course of describing the semantic interpretation of the syntax of a graphic display we are faced with describing how the marks map into semantic classes. Thus, it is at the point of formulating the semantic description that it is most convenient to begin to consider our two general mapping principles, derived from the requirements of notational systems (streamlined for present purposes). These principles deal with external, "vertical" mappings between levels, and internal, "horizontal" mappings between elements at the same level of description, and thus will sometimes be involved in the syntactic analysis per se.

The vertical mapping principle. Every meaningful difference in the value of a variable should be represented by detectable differences in marks, and every mark should have one and only one meaning. Ambiguous or missing marks violate this principle and require an alteration at the level of syntax.

The horizontal mapping principle. Portions of the chart or graph that are meant to correspond to other portions of the chart or graph should do so in an unambiguous way. The key, for example, should clearly indicate how labels are paired with different components of the specifier. This is true both at the level of the marks and at the level of the meaning of the marks (most notably labels). This principle is distinguished from the natural organizational principle in the following way: when a natural organizational principle has been violated, the violation can be corrected by rearranging marks already in the display (by repositioning lines and the like). When the horizontal mapping principle has been violated, new marks must be added (e.g., lines or arrows connecting parts). A necessary ingredient is missing when the mapping principle is violated.

Applying the Analytic Scheme

As in our treatment of the syntax of charts and graphs, we decompose the problem of describing the semantic content (the literal meaning) of a graphic display into four parts: characterizing the background, the framework, the specifier and the labels. As before, when describing the chart or graph, we are looking for violations of the operating principles that come to light when the display is being analyzed.

Background

If the background is patterned, the meaning of the pattern should be consistent with the information presented in the chart or graph. If background figures are present, do they distract from the meaning of the chart or graph? Are the elements of the background ambiguous? Is it clear whether elements are contiguous or overlapping? Do parts of the background occlude parts of the framework such that information is lost?

Framework

The most important feature of the framework is that it serves to allow the reader to extract the meaning of the marks and their organization. The elements of the framework should serve these ends. We begin by asking whether meanings of the elements are unambiguous. We note whether any part is not present or not implied. Next, we consider whether the syntactic properties of the elements engender correct mapping into a compliance class. Thus, we assess the scale type used in a graph and note whether the semantic scale is clearly indicted syntactically. For instance, if the scale used on the axes of a graph is syntactically dense, the semantics--the actual scale being represented--should also be semantically dense (e.g., a ratio scale should not be used in making the axes to represent an ordinal scale). In the same vein, the labels along the axes should be compatible with the actual scale being used and with the markings along the axes; the numbers spaced along the axis should suggest the correct scale type. Many of the problems with frameworks, as the reader probably inferred from the foregoing concerns, are violations of the formal mapping principles. The principles of surface compatibility also are sometimes violated here. Thus, we also ask whether variations in size, color, and the like are compatible with what is being represented (color changes from red to blue should not indicate rising temperature, for example).

We next note the extent of the scale, attending to not only its range, but the baseline. This may prove important in the subsequent analysis of the pragmatics of the chart or graph.

In addition to the foregoing questions, we check whether the lines that compose the framework depict some object. (This is quite common in many popular magazines). Thus, we ask:

> If the framework is serving to depict some object or scene, what is the meaning? Is the meaning clearly evident, and is the depicted object clearly representative of the class of objects being depicted?

The specifier

The meaning of the specifier is derived from how it relates parts of the framework together. Thus, in large part we will defer discussing the meaning of the specifier until considering the relationships among the different constituents. However, we can ask two things about the meaning of the specifier marks per se. First, they should be concise, no more or less being present than is needed to convey the information. If too little is present, the vertical mapping principle will be violated; if too much, it may be unclear what is being conveyed. (Note: if one wants a decorative piece of art accompanying an essay, however, this will be a violation only if the illustrations and fancy extraneous interfere with comprehension of the actual content.) Second, specifiers often are depictions (e.g., a graph of rising prices could have a jet plane taking off, with its exhaust being the function). If so, we ask:

> Are the depictions clearly representative of the compliance class in question? One would not want a picture of a potato to stand for "plant life," for example (since potatoes are hardly typical --in Rosch's (1978) sense--plants).
>
> In addition, one wants to ensure that marks used to represent different things look more different than marks used to represent the same thing. Further, one whould check that the literal interpretation of the marks is compatible with the role they play, as noted in our principles of surface compatibility.

Labels

For each type of label, we begin by considering whether the marks used as labels are compatible with the represented concept and whether the meaning of each label is accessible to the intended reader. Following this, more particular questions are asked of each of the three types of labels:

Alphabetic labels: Are the words ambiguous? Are the meanings of all the words representative of the class being indicated?

Numeric labels: Are the units clear? Are the units familiar?

Depictive labels: Are pictures used as labels easily identified; are they familiar to the intended readers? Are the marks used to depict clearly representative of the concept that they stand for?

Pair-wise combinations of labels: Cases where labels are serving to identify other labels (e.g., naming a picture) are also considered vis-a-vis our principles.

Organization of basic level constituents

Following analysis of each of the individual constituents, we again turn to an analysis of the organization of the constituents of the chart or graph. The way charts and graphs are organized is considerably more complex at the level of semantics than at the level of syntax, which also seems to be the case in language. We have devised two general kinds of rules of combination that are critical for deriving all of the information represented by marks in a graphic display. One kind of rule is appropriate for graphs, where a quantitative relationship between two or more values on two or more scales is represented; with two scales there are 10 possible combinations among the four scale types (nominal, ordinal, interval and ratio) that are commonly used. The other kind of semantic rules of combination is appropriate for charts, where a qualitative structure or organization of entities is represented. Let us consider each kind of combinatorial scheme in turn.

Quantitative Relational Information. Perhaps the best way to present the formal properties of this aspect of graphic semantics is in tabular form. Thus, the following table relates values on two scales to each other. We will consider all possible combinations of nominal, ordinal, interval and ratio scales except the nominal-nominal relations (which fall in the second class of rules). Recall that nominal scales are not ordered, with numbers being used as names (as on football players' sweaters); ordinal scales are rank ordered according to quantity, but the actual magnitudes of differences are irrelevant (as in the first, second and third place winners of a race); interval scales are ordered so that the magnitudes of differences mean something, but ratios of numbers do not (as in the Farenheit scale, in which the point labelled "zero" is completely arbitrary); finally, ratio scales have numbers that are ordered so that the magnitudes of differences are meaningful and ratios can be computed (as in Kelvin degrees, where 10° is twice as hot as 5° -- which is not true with Farenheit degrees). In addition to providing an example for each in the table, we list examples of the kinds of information available in each case. Extensions to n-dimensional cases follow in a straightforward manner from the simple two dimensional cases considered here.

INSERT TABLE 2.1 HERE

The information content of a graph can then be assessed by interpreting the individual axes, noting how points are paired by the specifier(s), and then and using the taxonomy in the table to derive the relationships between the values. If the relationship is not clear, there is a failure of internal mapping (the specifier is not clearly serving to pair points on the framework) or a failure of external mapping (part may be missing). (Violations of many other principles can also distort the relationship, depending on problems in seeing the specifier or organizing parts of it correctly.)

Structural/organization information: A computer flowchart, an organizational chart for a government agency, and a family tree do not relate values on dimensions. Rather, they specify the relationships among discrete members of some set. This sort of information can be described using the following three general criteria. These criteria are independent of one another.

The first criterion is whether the links between entities are directed or nondirected. Elements of the framework (i.e., marks indicating an individual member of the set) can be related together either by symmetrical or by asymmetrical relations. For example, in a kinship diagram, the vertical links of the tree are directed, indicating who is the parent of whom (an asymmetrical relation). The horizontal links, such as "sibling of" (a symmetrical relation), are nondirected.

The second criterion is how many types of links are used. More than one kind of relation may be used in a graph. In a kinship diagram, for example, "cousin of" and "brother of" may both be present. In a computer flowchart, only a single arrow--indicating which operation follows another--may occur following an operation.

The third criterion concerns the type of mapping used. There are three classes of mappings:

One:One, Many:One (or One:Many) and Many:Many mappings, which we will consider in turn: One:One mappings: In this case links in a chart might indicate how husband and wife pairings occur by drawing lines connecting points representing the location of each individual at a cocktail party.

Many:One or One:Many mappings: In this case, it is important to consider separately directed and nondirected links. With directed links, inclusion relations may be indicated by a Many:One mapping such as occurs in a hierarchy where many objects are organized under a superset. With nondirected links, collateral relations are indicated. If all diplomatic relations were symmetrical, links on a map illustrating the diplomatic relations of any one country would represent this sort of mapping.

Many:Many mappings: In this case, the multiple affiliations of a number of different objects can be represented. For example, a chart might represent different social classes by a drawing of a typical member of each, and might represent different social institutions by drawings of typical buildings (e.g., a church or a bank). Lines could connect the people to the institutions to which at least a majority of the represented class belong.

In charts, then, the nature of the mapping must be clearly indicated by the specifiers. Too many arrows can obscure mappings among elements, as sometimes happens in tangled organizational charts. Directionality and specific meaning (achieved via labels) may be important, and clearly defined links are always important. In actually describing a chart or graph, we are careful to consider what kind of information is being conveyed (hierarchial, relational, etc.). We then consider whether the marks effectively convey the meanings of the relations among the marks as the graph maker intended.

In the course of describing the overall organization among the constituents, we take special care on the following points:

Organization among the framework and labels. We consider how labels serve to interpret different aspects of the framework. Each label type is examined separately.

Organization among the labels and specifier. We consider how labels serve to interpret different aspects of the specifier. Each label type is examined separately.

Organization among the labels, framework, and specifier. Finally, we examine the overall configuration of the display, investigating whether graphic relations among depictions convey the intended meaning. We ask whether all associations among adjacent or overlapping material are clear, if the graph is not intact (perhaps because adjacent material on the page occludes part of it), and if it is difficult to read.

THE PRAGMATIC ANALYSIS

As in language, not all the information humans gather from charts and graphs is dictated by the literal interpretation of the marks on the page. If the number of war dead were indicated in a bar graph by increasingly higher piles of bodies of dead children, to take a grisly example, the reader would probably not simply register the literal information conveyed by the height of the column. Similarly, if one bar in a bar graph were printed in bright orange ink, and two others in dull gray, that bar would be hard to ignore. This "pragmatic" aspect of communication with charts and graphs has been discussed at some length by Huff (1954) in his classic book, How to Lie with Statistics. The operating principles offerred here were determined primarily by considering the kind and order of the description of the lines a person would build up, with pragmatic "connotations" arising from these descriptions. The principles were then tested by constructing demonstrations in which visual properties were manipulated to produce descriptions at the semantic level which emphasize some parts of the information at the expense of others, often to the point of being misleading.

Some of these principles have rather direct correspondences to similar principles underlying language (see Grice, 1967).

Operating Principles

Two classes of principles capture the relevant pragmatic uses of charts and graphs. The classes contain numerous individual principles, however, and thus we shall defer discussing them until chapter 4. The classes are:

A. Principles of invited inference

Although a chart or graph may not mislead on the semantic level, it may invite us to misread it anyway. This can be done in numerous ways: truncating scales so that small proportional differences appear larger; varying the type of scale used (linear vs. logarithmic, for example); using inferred 3-D properties of a display so that we see things as bigger than they are, and so on. Some of these principles are directly reflected by Huff's (1954) advice about how to lie with statistics.

B. Principles of contextual compatibility

Most graphic displays are embedded in a context, either in text or in an oral presentation. The context and the semantic interpretation of the display must be compatible or comprehension of the display will be impaired.

Applying the analytic scheme

We again consider first each of the four basic-level constituents, and then turn to questions about the organization among them. This analysis differs from the foregoing ones in an important respect: The syntactic analysis resulted in a rather rich description of the chart or graph itself. This was necessary because many of the elements of the syntax fed into the semantic properties, and, hence, we needed to have the chart or graph described in a way that would allow us to consider each of the semantic principles. At the level of the semantic analysis, there was much less description per se. And only some of the semantics of some aspects of the chart or graph are relevant for this later pragmatic analysis. The pragmatic analysis itself, then, produces very little in the way of description of the chart or graph. Rather, the existing description is now rich enough, from the level at which the thickness and color of the lines is noted to the level at which the elements are interpreted, such that we can simply ask questions that probe for violations of specific principles. Thus, this analysis consists entirely of questions, as indicated below. These are "leading questions" in that the answers reveal violations of the operating principles described above.

The Background

Does the background imply information not explicitly stated in the display (e.g., as might occur if the background was a photo of war dead)?

Are the implications of background material consistent with the message and the content?

The Framework

Does the form of the framework lead the reader to extract the intended message easily?

Is there a truncated axis? Does this emphasize small proportional differences in ways not intended by the graph maker? (Note: sometimes graphic displays make a point in part by emphasizing certain small differences; in some cases this may be misleading, in others, not.

Are scales distorted? Is this compatible with the point of the chart or graph?

Are the value markings indicated sufficient for intended purposes?

Are the marks used to represent a given element of such a form that they lead the reader to distort relative comparisons?

If the framework is also serving to depict, does the meaning of the depiction help or hinder understanding the content of the chart or graph?

The Specifier

Are some equivalent elements made to appear more important than others (by color, width of lines and so on)? Is this appropriate given the point of the chart or graph? Does it help or hinder understanding its meaning?

Are marks used to represent a given element of such a form that they lead one to distort relative comparisons?

If the specifier depicts information, does the meaning of the depiction help or hinder understanding the content of the display?

The Labels

Is the visual dominance and form of the elements of each of the labels consistent with the point being made?

Are some equivalent elements inappropriately made to appear more dominant than others (by varying color, weight, etc.)?

Are words consistent with the terminology of the text?

General Organization

Is the meaning implied by the text readily apparent in the chart or graph?

Does adjacent material on the page distract from or enhance the graph?

Does redundancy, if present, help or hinder understanding of the graph?

Is there a deliberate use of perceptual distortion (e.g., of areas)?

III. TWO EXAMPLES

In the final section of this chapter, we present two examples of how the analytic scheme is actually applied. In both examples, we indicate where a violation was discovered in the course of generating the description; violations are indicated by the word "VIOLATION" followed by the name of the principle violated and the reason that principle was considered to have been violated. Note again that not all violations will necessarily impair reading the chart or graph at the level of detail intended by the designer. Violations reveal difficulties in extracting all of the information potentially available in a display, but this may be far in excess of that required to use the display as intended.

The first display we analyze is a relatively simple bar graph, and the second is a very complex multiple framework chart. Both of these displays were taken from U.S. government documents, the first from a Department of Transportation manual and the second from a proposed scheme for labeling food products from the Department of Nutritional Sciences. In later chapters of this book we will not only discuss what is wrong with given charts and graphs, but we will discuss how best to correct their faults. Much of the information necessary to correct a given display will be provided in the detailed presentation of the various operating principles, as will examples of how these principles can be used to advantage or disadvantage in preparing charts and graphs.

I. Analysis of Figure 2.3

INSERT FIGURE 2.3 HERE

SYNTACTIC ANALYSIS

The following description is for the graph illustrated in Figure 2.3. Note that if a question in the descriptive scheme is clearly inappropriate (e.g., about color when only black and white are used), it is ignored. Similarly, questions designed for special purpose problems, such as the relations among 2D and 3D depictions, are ignored if the graph includes only 2D information, as does this one.

Background

Blank white.

Framework

There is an outer and inner framework.

The outer framework

Elements: 2 vertical straight lines syntactically dense.
2 horizontal straight lines, syntactically dense.
Medium weight, black.

Organization: Connected to form a rectangle, with the vertical axis being longer.

The inner framework

Elements: 7 straight vertical lines, syntactically dense.

Organization: spaced evenly.

Organization of inner and outer frameworks

Inner lines connected to horizontal lines of outer framework, terminate at those lines.

The specifier

Elements: 5 rectangles, divided into black and white portions by a vertical line, with the left side being black; or, 5 black rectangles and 5 white rectangles.

VIOLATION: Principle of Processing Priorities. The width of the bars is visually dominant, which is distracting because the width has no information value.

Organization: Spaced one above the other with the leftmost ends aligned or, the black rectangles juxtaposed to the white ones, with the rightmost end of the black ones abutting the leftmost end of the white ones, and the pairs of rectangles being spaced vertically, with the leftmost ends of the black rectangles being aligned.

Labels

Title: Two fonts are used: Above a large label is a smaller one, part of which is a number.

Key: There is a key; analysis of it is presented in relation to other components below.

VIOLATION: As is evident below, the Internal Mapping Principle is violated. It is not clear if the white rectangle in the key corresponds to only the white part of the pictorial material (bars) or the entire bar.

Alphabetic

Two sizes of typefonts are used, they will be referred to as either "large" or "small."

VIOLATION: The Principle of Processing Priorities. The size of the letters labeling the two scales is varied arbitrarily, making one more salient for no good reason.

Vertical axis label, small font
Key labels, small font
Total distance label, small font
Horizontal axis label, small font

Organization: In relation to framework, as noted below.

Numeric

Five in vertical column on left, large font
Five in vertical column at right, large font
Six in horizontal row at bottom, small font
Number in title at top, large font

Organization: Rows and columns, right and left columns in 1:1 correspondence; in relation to framework and specifier, as noted below.

Depictive

Key: Black rectangle, white rectangle

Organization: Adjacent to each other, and in relation to both framework and alpha labels, as noted below.

Organization among the different types of labels

Alphabetic and Numeric

Left: vertical line label above column of numbers.
Right: total distance above column of numbers.
Bottom: line label to left of row of numbers.

VIOLATION: Gestalt Principle of Organization (similarity). The size of the marks used as labels on the vertical axis and the size of the marks used as numbers are incompatible, making it difficult to see them grouped together.

Numeric Depictive

No cases.

Alphabetic and Depictive

Labels to right of white and black bars.

VIOLATION: Gestalt Principle of Organization (proximity). The MPH label is not clearly associated with the vertical scale, being in a non-conventional location.

Organization among the framework, specifier and labels

Framework and specifier

Bars abut left vertical line with bars extending to right.

Bars enclosed in frame.

Vertical internal lines of frame do not violate boundaries of rectangles.

Framework and labels

Alphabetic

Title at absolute top, key directly above highest horizontal line of outer framework.

Labels of left vertical straight outer line and bottom line outside framework. Label at top left, at bottom with first letter directly under extreme point of bottom horizontal line.

Total distance label at upper right above top horizontal line, centered within segment defined by first internal vertical line to the left of the right outer line of framework and the right outer line.

Numeric

Column on left regularly spaced outside and to left of leftmost vertical line of outer framework.

Row on bottom under horizontal lower line of outer framework, one number under each internal line, no number under last internal line on the right. Column on right, evenly spaced, centered between first internal line to left and rightmost outer line.

Depictive

Above horizontal line defining top of framework.

Labels and specifier

Alpha

No cases.

Numeric

1:1 alignment of right column of numbers and bars.

VIOLATION: Gestalt principle of Organization (proximity, similarity, continuity). Numbers are not clearly grouped perceptually with appropriate bars.

Depictive

Black and white key labels in same order as black and white portions of bars. Not clear of white box corresponds only to white portion of bars.

VIOLATION: The Internal Mapping Principle, as noted in the initial comments on the key.

Labels, framework, and specifier

Description of the pair-wise relations among the constituents is sufficient; no special problems emerge from the constituents taken as a whole.

SEMANTIC ANALYSIS

We have now described the basic elements on the page and their organization at a level sufficient to consider how these marks act as symbols. Let us again consider each aspect in turn. We will first begin by considering the interpretation of the syntactic units just described.

Framework

Outer

A Cartesian coordinate space is defined by the horizontal and vertical lines.
The vertical axis represents a ratio scale, with the origin at the top of the line. Although this scale is semantically dense, it has been differentiated into five descrete values with values increasing as one descends down the line.

VIOLATION: Principle of Graph Schema Availability. The vertical scale violates a common graph form, in which larger values are usually indicated by higher marks. The origin of the two axes in a Cartesian space is usually the same point (the lower left intersection of the axes), which is not true here.

The horizontal axis is a ratio scale, with the origin at the left and values increasing as one moves to the right.

Inner

The vertical lines mark off increments of distances of 50 feet.

The specifier

Length of the entire rectangle represents average braking distance.

Length of the black portion represents average reaction distance.

Each rectangle represents a discrete and different speed.

The relationship between average braking and reaction distance is implicit in the relationship between the length of the black and white portions of the bars.

VIOLATION: External Mapping Principle. The ambiguity in how to describe the specifier on a syntactic level violates the requirements of our streamlined version of Goodman's concept of a notational system, as described earlier.

Labels

Elements:

Alpha

English words labeling the values of units on the axis, the meaning of the depictive label used in the key, and the meaning of the total distance column. English words also label the graph as a whole.

VIOLATION: The External Mapping Principle. The failure to include the word "distance" on the alpha label associated with the left bar in the key is misleading as no contrast is intended to the right label.

Numeric

Distances in feet and speed in miles per hour. Also total braking distance. The figure is related to textual material by a number at the top.

Depictive

Color of bars in the key have no intrinsic meaning.

Organization:

Alpha and Numeric

Words label scales that the numbers index values on.

Alpha and Depictive

Words label the meaning of the bars in the key via a 1:1 mapping.

Numeric and Depictive

No cases.

Relationships among the framework, specifier and labels

Having described the interpretations of the units defined syntactically, let us now consider the interpretation of the relationships among these components.

Framework and specifier

Outer framework

The specifier is serving to map discrete values on the vertical axis to continuous values on the horizontal one (although both are ratio scales).

Two functions are plotted, and the relationships between these two functions can be computed.

Framework and labels

Alpha

The labels define the meaning of the axes.

Numeric

The numbers on the vertical axis serve to differentiate the ratio scale into five descrete classes.

The numbers on the horizontal axis demarcate values on a dense ratio scale.

Depictive

The bars in the key label and the bars in the framework via a one:many map.

Framework, specifier material and labels

The semantic relations are described in the quantitative semantics in a straightforward way, as is evident in the descriptions given for the pair-wise organization among constituents.

PRAGMATIC ANALYSIS

There are no violations of pragmatic principles evident; seeing the graph in context could reveal some, but we will not consider any such context here.

II. Analysis of Figure 2.4

INSERT FIGURE 2.4 HERE

SYNTACTIC ANALYSIS

Analysis into subcharts

The chart is divided into two subcharts (left and middle) and a cluster of alpha and numeric material (hereafter referred to as the right table). The rightmost boundary of the left chart is defined by right justification of seven circles and blank space to the right of the circles. The rightmost boundary of the center chart is defined by annular white space between the small radial marks in the center of the page and the circular justification of the alpha material on the right.

Left Subchart (LS)

LS Framework

The LS framework consists of an outer frame and an inner frame.

The outer framework

Elements: A horizontal axis is indicated by the bracket on the bottom. Axis is syntactically differentiated.

Organization: Only one element in outer framework.

The inner framework

Elements: Twenty-eight (28) closed curved lines, forming circles. These are syntactically dense. Medium weight, black.

Organization: Circles aligned into columns via proximity.

VIOLATION: Gestalt Principle of Organization (proximity). Proximity results in an organization into columns when an organization into rows is required.

Organization of inner and outer frameworks

Bracket encompasses inner framework elements.

LS Specifier

Elements: Black quadrants of circles (i.e., subtending 90° of arc).

Organization: Contained within LS inner framework elements. When one of these elements appears in a frame, it is positioned in the upper left quadrant. As additional elements are added to a frame, they are placed contiguous to prior elements and fill the frame in a counter-clockwise manner. Frames are filled from left to right in rows.

VIOLATION: Gestalt Principle of Organization (good form). At first glance, the inner framework leads one to divide the quantities into fourths, which is incorrect.

LS Labels

Only alpha and numeric labels appear - there are no depictive labels. Since alphas and numerics appear in the same perceptual units, separate syntactic discussions seem inappropriate.

One typefont (medium weight, black) is used within this subgraph and alphas may be upper or lower case.

Elements: Subchart title - "Nutritional cont." The first letter is upper case, remaining letters are lower case, a period appears last.

Organization: Letters have upright orientation and are arranged in two groups in a closely packed horizontal string.

Elements: Seven vertical axis (row) labels are mixed upper and lower case with periods and numerics intermixed.

Organization: Letters have upright orientation and are arranged in one or two groups in closely packed horizontal strings. Labels are left justified at the same column.

Elements: Horizontal axis label - "needed per day" is composed of lowercase letters.

Organization: Letters have upright orientation and arranged in three groups in a closely packed horizontal string.

Organization Among Different LS Label Elements

LS title is left justified in the same column as the vertical axis labels. The space left between the title and the top vertical axis label is only slightly greater than the space between the various vertical axis labels.

Organization Among the Framework, Specifier and Labels

Framework and Specifier

The dark quadrants of circles are contained within inner frame elements, as mentioned above.

Framework and Labels

The title is just above and commences to the left of the array of circles.

The horizontal axis label is below the bracket.

VIOLATION: Gestalt Principle of Organization (proximity, similarity). Both the position of and use of the same typefont for all labels impairs identifying the subtitle as distinct.

Labels and Specifier

The specifier is not labeled.

Middle Subchart (MS)

MS Framework

The outer framework

Elements: 40 short lines, approximately equal in length. The frame compromised of these elements is syntactically differentiated.

Organization: The lines project outward from a common center and extend from a common distance from the center and to a slightly greater common distance from center. The lines are separated by approximately equal angles, but the separating angles are discriminably different.

VIOLATION: Principle of Processing Limitations. That there are exactly 40 short marks in this frame is not immediately apparent, but is important in order to understand the chart.

The inner framework

None

MS Specifier

Elements: Two "pie-slice" wedges; one black, one white. The black is slightly larger than the white.

VIOLATION: Gestalt Principle of Organization (good form). Failure to include the rim of the white wedge impairs seeing it as a wedge.

Organization: The curved edges of the wedges are conterminous with the distal end of the frame elements. The vertex of the black wedge points straightdown while the vertex of the white wedge appears to point straight up. The vertices are joined.

MS Labels

No labels are present within the subgraph.

Organization Among Different MS Labels

Framework and Specifier

Both wedges have vertices which coincide with the center of the circle defined by the frame. Both wedges obscure the short radial lines which define the frame.

Framework and Labels

Not applicable.

Labels and Specifier

Not applicable.

Right Table (RT)

RT Framework

No explicit framework, outer or inner.

RT Specifier

There is no specifier in this table.

RT Labels

There are both alphabetic and numeric labels in this table. No depictive elements appear. Two typefonts are used: One is small light upper case, the other is large bold lower case. All letters and numbers in the same cluster have the same typefont. Right justification is apparent for entire table, with the exception of the digit "8".

VIOLATION: Gestalt Principle of Organization (proximity, good continuation). The "8" being out of line in the top cluster leads one to focus one's attention on it, for no good reason.

Alpha

Elements: Three rows of small, upper case type are at the top. Spacing divides these rows into two columns. Alphas appear in only one (top) string of right column. (Numeric "8" is also in right column). Beneath these are three more rows, bold type, in lower case. Spacing again produces two columns. Beneath these elements is one row in bold lower case type.

Organization: Typefont and weight of lines serve to define three groups, as noted above. The top group is organized into a row of one line and a row of two lines (by the Gestalt Law of Proximity). The middle group is directly beneath the first, being aligned on the right margin. The final line is separated from the rest of the table by a large gap.

VIOLATION: Gestalt Principle of Organization (proximity). The large gap separating the bottom line of the table impairs one realizing that it belongs to the table.

VIOLATION: Principle of Processing Priorities. The difference in font size between the upper and middle clusters direct one's attention to the middle cluster first, instead of the top one.

Numeric

Elements: Numerics appear in each cluster.

Organization: When more than one numeral appears in a string, they follow one another in sequence. They appear in the right-most perceptual units of the table, except in the bottom line.

RT Organization Among Different Types of Labels

Alphabetic and Numeric

Numerics, when present, are intermixed in the same perceptual units with alphabetics.

Macro-Organization

Having discussed the syntax of the various subcharts, we return to overall structure of the three.

Framework

Elements: Two heavy black lines composed of a short vertical segment and a longer horizontal segment ending in an arrowhead.

Organization: One line originates at the center of the rim of the black wedge and terminates at the left in an arrowhead, which points at the right-most part of the title of the left subchart. The lower line originates at the center of the rim of the white wedge and points at the left-most end of the bottom line of the right table.

Labels of Macroframework

Only alpha and numeric labels appear - there are no depictive labels. Since alphas and numerics appear in the same perceptual units, separate syntactic discussions seem inappropriate.

Elements: Title - the title is comprised of two perceptual units; one in small upper case, one in very large upper case.

Organization: Upright orientation, arranged in horizontal strings comprised of two or more closely packed groups. Small typefont is centered in the page and above very large typefont.

VIOLATION: Gestalt Principle of Organization (similarity, proximity). The title is not clearly identified as such. It should be either set off from the chart proper and/or be in a heavier typefont.

Organization of Macroframework and labels

Both perceptual elements of the title are centered above the framework.

Overall Organization

VIOLATION: Principle of Processing Limitations. There is too much information to process at once.

SEMANTIC ANALYSIS

Left Subchart

Framework

Outer framework

Vertical axis (implied by white space to the left of the left most column of circles) constitutes a nominal scale. This scale is semantically differentiated (although differentation is de-emphasized perceptually by wider spacing row-wise than column-wise, as noted earlier). Horizontal axis constitutes a ratio scale and is semantically differentiated. The extent of this scale represent daily nutritional requirement of given nutrients. The bracket functions as a way of indicating the scope of the label on the bottom, as will be discussed shortly.

Inner framework

Each circle in a row may contain as much as 1/4 of the daily requirement for a given nutrient. The circles are thus ratio scales and are semantically differentiated.

VIOLATION: External Mapping Principle. The semantic differentiation is made apparent only through the relationship of the specifier with the inner framework. The perceptual representation of these circles actually falsely suggests a dense scale by the lack of differentiation marks on the circle.

The specifier

The basic specifier unit (a black quadrant of a circle) represents 1/16 of the daily requirement for a given nutrient. Basic specifier units can be combined to indicate integral multiples of 1/16 of the daily requirement.

VIOLATION: Principle of Schema Availability. The nesting of quadrants within each of the four circles is a novel way of specifying the information, and hence, must be clearly specified.

Labels

Elements:

Alphabetic

English words are used in the title to inform the reader that the subchart provides information on nutritional contents. They are also used to name the various nutritional components represented as rows of circles and to inform the reader of the meaning of the horizontal axis.

Periods (.) inform the reader that a sequence of letters is an abbreviation of an english word.

Numeric

Numerals appear as characters which, in part, form the names of the nutritional components.

Organization Among Different Types of Labels

Alphabetic and Numeric

Together comprise names.

Relationships Among the Framework, Specifier, and Labels

Outer framework and inner framework

The bracket can be interpreted as unifying the collection of four circles into one dimension (along the horizontal axis of the inner framework). This is a One:Many mapping.

Framework and Specifier

The basic specifier units (black quadrants of a circle set in conjunction with the four circles in each row to indicate the extent to which one serving of the food item satisfies the daily requirement for a nutritional component associated with the row.

Framework and Labels

Alphabetic

The alpha labels define the meaning of the axis. The bracket indicates that the horizontal axis is defined by the English words immediately beneath it. This is a One:One mapping.

Numeric

Act in concert with alpha to name nutritional components represented by rows.

Framework and Specifier

The specifier is not labeled directly.

Middle Subchart

Framework

This framework is ambiguous. The only interpretation that is consistent with the other subcharts in the display is that this one framework represents two distinct entities. One entity (the top part) is the total daily nutritional requirement for a person. The second (the bottom part) is the total daily caloric requirements for a person.

Accepting these interpretations, the framework(s) would constitute a ratio scale.

While the frame appears syntactically differentiated, on the semantic level, the issue of denseness and differentation appears completely indeterminate in the context of all information present or derivable.

VIOLATION: External Mapping Principle. The ambiguity mentioned above is due to faulty mapping from syntax to semantics.

VIOLATION: External Mapping Principle. The variation in spacing between the marks of the frame seems to have no meaning.

The Specifier

The black wedge represents the proportion of the total daily nutritional requirements supplied by a serving of the food in question (This interpretation is the only one consistent with the connective relation between the black wedge and the left subgraph.)

The white wedge represents the proportion of the total daily caloric requirements supplied by a serving of the food in question. This interpretation is uncertain, however, but is suggested by the fact that the arrow from it points to the bottom line of the table on the right.

VIOLATION: External Mapping Principle. The meaning of the wedge simply is not clearly defined on the syntax or the semantic context, allowing one to interpret the meaning of the syntax in more than one way.

VIOLATION: Principle of Graph Schema Availability. A circle or "pie" chart is usually used to show how a whole is divided into parts. The middle subchart, on the other hand, does not use wedges to divide a single entity into parts, but rather treats the two wedges as independent.

Labels

No labels of any sort are wholly *within* subchart.

VIOLATION: External Mapping Principle. Missing labels on both the framework and the specifier make this chart very difficult to understand.

Relationship Between the Framework and Specifier

According to the most consistent reading, the specifier elements represent two distinct entities: (1) proportion of daily nutritional requirement supplied per serving (black wedge), and (2) proportion of daily caloric requirement supplied by a serving (white wedge). The frame represents the whole daily requirement of these two entities (nutrition and calories) and, therefore, supplies ratio scales in which both specifier elements are measured. The different sizes of the two wedges is thus explained.

VIOLATION: External Mapping Principle. If this interpretation is correct, the scale is different things to different objects, and therefore, violates the disjointness property required for systems of symbolic notation to be unambiguous.

VIOLATION: Internal Mapping Principle. The wedge-shaped specifier elements obscure the hash marks which comprise the outer framework. This prevents any quantitative mapping from specifier to frame.

Framework and Labels

The frame is not labeled in this subchart. If it had been, two different labels would have been required for the same framework or the framework would have to be divided into two semicircular frameworks, each separately labeled.

Labels and Specifier

The specifier in this subchart is not labeled within the subchart. Specifier elements within the subchart are connected to labels in other subgraphs and derive meanings thereby, as will be discussed shortly.

Right Table (RT)

RT Framework

There is no actual framework.

RT Specifier

There is no specifier.

RT Labels

Alphabetic

The labels in the upper cluster are English words which specify quantities of food. The labels in the middle cluster are English words for abbreviations which are names of nutritional components of food. The symbol "g" indicates "grams."

The lower label is an English word meaning a unit of heat (in this context, the heat equivalent of a serving of food).

Numeric

The numerics are arabic numerals specifying quantities.

RT Organization Among Different Types of Labels

Alphabetics and Numerics

Alphabetics and numerics appearing in the same perceptual units together specify a quantity of some type of physical units (e.g., "4 grams"). These units in turn specify how much of the named substance associated with the quantity in a serving.

Macro-Organization

Framework

One arrow associates the white wedge with the "170 kilocalories" label. This, in fact, allowed us to infer the meaning of the white wedge.

The other arrow associates the black wedge with the entire left-most subchart, which provides an analysis of the total daily requirement of the nutritional components.

VIOLATION: External Mapping Principle. The lack of labels on the arrows impairs one from realizing that they symbolize different relations, "decomposes into" (top) and "corresponds to" (bottom).

VIOLATION: Principle of Graph Schema Availability. Arrows point from specifier elements to labels in place of the more conventional directions from label to specifier elements.

Labels of Macro-framework

Elements:

Alpha and Numeric

The title identifies this display as the third in this chapter, and labels the information provided by the entire set of charts.

Overall Organization

VIOLATION: Internal Mapping Principle. Labels are missing that are necessary to coordinate the subcharts into a single cohesive display.

VIOLATION: Internal Mapping Principle. One cannot easily relate the information about protein in the right table to the information about protein in the left chart, partly because of the use of "prot." and "protein" in the different subcharts. In general, use of different notations or abbreviations lends one to infer that different things are being talked about.

VIOLATION: Principla of Internal Mapping. One must realize that there are forty marks comprising the frame elements in order to construe a consistent relation between the left and middle subgraph (in terms of nutritional content). The marks should have been emphasized (e.g., every tenth made bolder) to facilitate this realization.

PRAGMATIC ANALYSIS

There are no clear cases where the display has been slanted to lead us to draw incorrect inferences or attend to specific pieces of information more than others. We cannot know whether the pragmatic principle of contextual compatibility is violated because we do not know the context in which the display occurred.

CHAPTER 3: SYNTACTIC PRINCIPLES

I. Seeing the lines
  1. Adequate discriminability
    a) Relative distinctions
    b) Detecting marks
  2. Perceptual distortion
    a) Optical illusions
    b) Systematic distortion

II. Natural units
  1. Gestalt laws of organization
    a) Good continuity
    b) Proximity
    c) Similarity
    d) Good form
  2. Integral/separable dimensions

III. Processing priorities and limitations
  1. Priorities: salience
    a) weight and noticeability
  2. Limitations: fixed capacity
    a) 7 ± 2: "finite capacity;"
    b) Comparing units or parts thereof

CHAPTER 3: SYNTACTIC PRINCIPLES OF CHART AND GRAPH CONSTRUCTION

In this chapter we begin to consider principles that must be obeyed if a chart or graph is to be readily comprehensible. The principles specifically addressed in the present chapter concern how lines on a page are seen, organized, and held in mind. In the next chapter we will consider how such patterns are interpreted as meaningful units and how conceptual and quantitative information is extracted from them. In both this and the following chapter, each of the principles we present is illustrated by at least one "before and after" pair of displays, demonstrating how a violation of the principle clearly impairs graph reading, and how such violations can be repaired, thereby improving graph reading. Thus you, the reader, are in a sense a subject in an informal experiment: if you clearly agree that our repair of the "before" graph improves its legibility in your eyes, we may take it as prima facie empirical support for the validity of the relevant principle. This methodology has been employed successfully in the study of linguistics and in the study of perceptual illusions, constancies, and organizing principles. In addition to these demonstrations, in the sections to follow we summarize the available empirical findings--in the literature at large and on charts and graphs in particular--that bear on each principle and we present new data bearing on each principle.

The syntactic operating principles all rest on facts about how we see and encode visual information. Thus, the support for these principles is of two kinds, direct investigations of charts and graphs per se and more general studies of human visual information processing. The relative paucity of research on charts and graphs is more than balanced by the richness of our knowledge about visual perception. Hence, we are in a position to formulate the syntactic principles with a high degree of confidence. In each case we can not only marshall evidence that the principle is correct, but provide details about how to avoid violating the principle and how to make use of it in effective presentation of information in graphic displays.

In the remainder of this chapter we will consider three general classes of principles. The first class of principles must not be violated if the lines on a page are to be seen correctly. These principles deal with the acuity of the visual system and with the way in which the lower levels of the visual system systematically distort the simple attributes of what we see. The second class of principles specify the factors that determine how we group marks into units. These grouping principles are especially important because they determine whether the basic-level graphic constituents and relations among them (e.g., which part of the display is labeled by a given word) will be detected easily. The third class of principles outline factors that determine the priorities and limitations of visual processing. These last principles deal not with perceptual processes per se, but rather with the process of encoding information into memory. In particular, we consider the limits of "short-term memory", which place real constraints on how many units a graph maker can sensibly expect a reader to process at one time.

Thus, in this chapter we trace the path of visual processing of a graphic display, beginning with very low-level physically-defined attributes and ending with attributes that are fairly removed from the eye and visual system per se and more closely linked with abstract conceptual thought. This path, from outside to inside, will be further charted in the ensuing chapter when we leave the realm of perceptual processing altogether and consider the linguistic and conceptual underpinings of graphic comprehension.

I. Seeing the Lines

Two general principles codify factors that affect how well we see the lines that comprise a graphic display. The principle of adequate discriminability specifies the size of the difference between two marks that is necessary for us to detect it, and how pronounced a mark must be to be seen at all. The principle of perceptual distortion specifies how the visual system systematically distorts some visual dimensions, leading us to make increasingly larger errors when comparing marks of larger magnitudes.

1. The Principle of Adequate Discriminability

There are four different ways in which visible marks can vary, and associated with each are many different dimensions that potentially may be used to code information. First, a mark may vary in its quality. For example, differences in color or visual texture of a particular mark can convey information. Similarly, the position of a mark on the page may be informative. Second, a mark may vary in intensity. Brightness, lightness, and density or numerosity are dimensions along which intensity of a mark may vary. Third, a mark may vary in its extension, such as its length, area or volume. Finally, a mark may vary in duration, which may be important in dynamic displays such as Traffic Situation Displays (Warner, 1969) and so-called "kinostatic" or time-varying graphs discussed in Biderman (1971) and Wainer and Thissen (1981).

If a mark is to map uniquely into its corresponding "compliance class" at the semantic level of description, variations along any dimension must be perceptually different. That is, the reader must be able to detect differences in magnitudes of information-conveying marks. Thus, good graph making will be aided by data on human abilities to detect and discriminate variations along the physical marks. For example, data on the smallest point or difference in length that a person can detect under normal viewing conditions will help ensure that displays are legible, especially if the original display is reduced in size for publication. In this case, the data define limits in our ability to make absolute discriminations, to detect the presence of a mark. Similarly, if a comparative judgment is to be made of differences among marks, then data on minimum perceptible differences are necessary to ensure that there will be no ambiguity in difference judgments, including cases in which the graphic display is reproduced at different sizes. This corresponds to limits in our ability to make relative discriminations.

The limits in our abilities to make discriminatory are a consequence of the nature of our perceptual system. For example, before the physical mark can even affect the sensory receptors of the eye (the "rods" and "cones"; see Kling and Riggs, 1971), it must be projected onto the retina. This projection is accomplished by the refractory properties of the lens of the eye and changes in these refractory properties caused by accommodation. Because of factors such as optical defects, deviation from sphericity of the refracting surface, scattering, and wavelength-dependent properties (Field and Magoun, 1959), the quality of the retinal image is necessarily degraded, limiting the resolving power of the visual system as a whole. This degraded retinal image is transduced by the retinal cells into a frequency code of all-or-none action potentials which are then transmitted via the optic nerve lateral geniculate nucleus of the thalamus, and optic radiations to the visual cortex for further processing. Anatomical and physiological properties of the receptors themselves (Abramov and Gordon 1973) and mechanisms of neural transmission and decoding (Aidley, 1971) contribute to further limitations in our ability to detect and discriminate variations in the physical properties marks.

Neurophysiological phenomena allow us to explain, in part, some of the reasons for finite discrimination, but because of their complexity and our limited understanding of the mechanisms involved, we cannot yet use them to explain all the perceptual data. It thus becomes necessary to analyze performance at the level of the entire visual system, especially because the behavior of the system as a whole is of prime concern here. One way of proceeding at this level is to treat the human as a measuring instrument for visual inputs and to describe the performance of the input-output behavior of this instrument. For absolute and relative discrimination tasks, the inputs are marks varying along any dimension, and the output is the response of the individual to questions about the presence or absence, difference or sameness, of the marks. Data are then obtained by varying the magnitudes along particular dimensions and noting the minimum variation that elicits a qualitatively different response.

In the remainder of this section we consider these two topics, absolute and relative discriminations. For each topic, we briefly discuss the concept of threshold and how it may be measured. Then we present data on thresholds for various physical dimensions, as well as contextual factors which influence these thresholds. At the same time, we present examples illustrating how the data may be exploited in designing unambiguous graphs.

Absolute Discrimination

Threshold Determination

The relevant research on our ability to make absolute discrimination hinges on the notion that there exists a fixed sensation magnitude, or threshold, below which a stimulus is never detected (sensed), and above which a stimulus is always detected. If a series of stimuli are presented with magnitudes near the threshold, there should be a well-defined separation of those stimuli that are sensed and those that are not sensed. The point which divides stimulus magnitudes into those which are "sensed" and those "not sensed" is called the absolute threshold.

Although the absolute threshold is theoretically fixed at some point on the stimulus magnitude continuum, measuring such a threshold is in no way a trivial task. The measuring instrument is a human subject whose response is not completely predictable or reliable. The effect of this is that the threshold is obscured by the "noise-producing" variability, and statistical measures must be used to extract the actual threshold. We will not discuss here the actual procedures used to compute the thresholds; the interested reader is referred to Luce and Galanter (1963). Furthermore, the very assumptions about the existence of a fixed threshold and the proper way of measuring it have been called into question, and modern researchers use the more sophisticated assumptions and techniques of the Theory of Signal Detection, which assigns a central role to the inherent statistical variability of the visual system and to the biases and motivations of the perceiver (Green and Swets, 1966). However, for our purposes, which are to glean rough estimates of the resolving power of the visual system for use in the design of readable charts and graphs, we may innocuously adopt the "classical" assumptions about sensory thresholds.

The two most important thresholds for graph construction are *visual acuity* and *contrast*. Data on acuity and contrast are important when considering legibility of labels and pictorial material, as is described below.

Visual Acuity

Maximum visual acuity is defined as "the smallest visual detail that we are capable of resolving at a specified distance." Visual acuity is expressed by the visual angle in minutes of arc subtended by the physical stimulus S. For a rough sense of the measure, hold out your thumb at arm's length; it subtends around 2° of visual arc. There are 60 minutes per degree of arc. For small angles, visual angle (in minutes of arc) is computed using the following formula:

$$\beta = (57.3)\,(60)\,\frac{S}{D} \qquad (1)$$

where S is the physical size of the mark referred to as the distal size and D is the distance from the eye to the mark. The constants in the formula convert the units of visual angle from radians to minutes of arc.

For present purposes, it is critical to note that detectability does not ensure legibility. Identifying a mark as being of a particular type is more difficult than merely noticing that some figure is present. The literature on legibility is well documented in the Human Factors literature (see Smith, 1979), and standards such as Military Standard 1472B (1974), established by the Department of Defense, are routinely available. Table 3.1 sumarizes some of the recommendations for sizes of display letters. As a rule of thumb, under normal viewing conditions one can assume a standard acuity of one minute of arc (Thomas, 1975). Given this specification, in order to recognize the details of the capital letter "E", for example, its vertical size would have to subtend at least two minutes of arc, one for each pair of its horizontal strokes. However, a "standard" acuity of one minute of arc corresponds to a detection probability of only seventy-five percent. If near-certain detection is wanted, 1.6 minutes of arc should be specified, making our letter "E" subtend about four minutes of arc. This corresponds to 0.021" seen from 18" away.

INSERT TABLE 3.1 HERE

Let us consider an example of how we might use the data on visual acuity to specify the type font necessary to ensure adequate legibility of the label ANGLE for a display reduced by a factor of 2:1. If we assume a normal viewing distance of 18", then a 2:1 reduction results in an equivalent viewing distance of 36". We previously determined that recognition of the letter "E" required a minimum of 4 minutes of arc. Therefore, at a viewing distance of 36" the required type font must be 0.17" using equation (1). And in fact, research on reading has shown that character sizes should be between .06" and .17" for maximum legibility (Spencer, 1969).

As an illustration of how the principle of adequate discriminability can be violated by a graph, consider the set of graphic displays shown in Figure 3.1 taken from an article by Wickens and Kessel (1977). At this level of reduction, the labels "Hits" and "Misses" associated with the key subtend a visual angle of approximately 4.5 minutes of arc. But to identify the letters correctly 75% of the time the visual angle must subtend at least five minutes of arc. Thus, these labels begin to violate the boundaries of our identification abilities and at a normal viewing distance of 18" the reader will notice that it does take some effort to make accurate identification. Compare this to the improved version on the right; this should be much less work to read.

INSERT FIGURE 3.1 HERE

Luminance

The trend to computer graphics has led us to consider luminance as an important contextual parameter affecting acuity. Luminance is the amount of light per unit area reflected from or emitted by a surface (this measure is frequently referred to as brightness, although brightness is the subjective sensation to changes in the physical energy of light). Luminance is expressed in a variety of units for which conversion factors are given in Table 3.2. The three preferred units of luminance are the Lambert, Millilambert and the Foot-Lambert.

INSERT TABLE 3.2 HERE

The Lambert (L) is defined as the unit of luminance equal to that of a perfectly diffusing and reflecting surface illuminated by a standard candle at a distance of one centimeter (cm). The Millilambert (ml) is one thousandth of a Lambert. The Foot-Lambert (ft-L) is defined as the unit of luminance equal to that of a perfectly diffusing and reflecting surface illuminated by one foot-candle. Normal reading light is about 10 ft-L. The luminance values experienced in a number of common situations are given in Figure 3.2.

INSERT FIGURE 3.2 HERE

Note from the Figure that as we move from low to high luminance levels, we move from "rod" to "cone" vision. Rods and cones, the two types of photoreceptors found in the eye, differ importantly in their spatial distribution and functional properties. Basically, cones provide acute vision during daytime luminance levels, whereas rod vision is most sensitive to low luminance levels and is essential for night vision. These and other important function differences are summarized in Table 3.3.

INSERT TABLE 3.3 HERE

Visual acuity is highly dependent upon the background luminance on which a dark detail is superimposed. Figure 3.3, taken from a study by Moon and Spencer (1944), shows the relationship between acuity and background luminance. As luminance increases, acuity increases--partly because the cones become active and, as Table 3.3 indicates, the spatial resolution of cones is much greater than that for rods. For normal reading light (about 0.1 ft-L), the eye can detect an object subtending about 1 minute of visual angle.

INSERT FIGURE 3.3 HERE

Contrast

A second factor that must be considered if a display is to be legible is our ability to discriminate displayed detail from visual background or to discriminate contrast in brightness. For details darker than their background (commonly the case for graphic displays), contrast can vary from 100 percent positive, to zero.

Contrast is a measure of difference in luminance between a detail (LD) and its background (LB) and is computed by the formula:

$$\text{Contrast (\%)} = \frac{LB - LD}{(LB)} \times 100 \qquad (2)$$

One empirical approach for determining the limits of this ability is to determine the minimum contrast needed to perceive a particular pattern. The simplest type of pattern is a grating made up of a series of light and dark bars. If the luminance difference between the light and dark bars is reduced sufficiently, there will be a point at which they are just discriminable. The point is called the contrast threshold; the lower this threshold, the greater the contrast sensitivity.

Our visual systems do not have a single contrast threshold for all stimuli. Rather, our contrast sensitivity differs depending on the sharpness or gradualness of a luminance change, being highest for intermediate degrees of gradualness and lower for extremely gradual changes and for extremely sharp changes (i.e., fine details). This relationship was discovered by observing the contrast threshold for grating patterns of different degrees of fineness. If the fineness of the grating is expressed as cycles per degree (number of light dark pairs subtending 1° of visual angle), then for gratings of any fineness, the contrast can be varied to yield the contrast sensitivity. A plot of this sensitivity for gratings of different spatial frequencies (fineness) can then be obtained and is referred to as the contrast sensitivity function (Campbell and Robson, 1968). Campbell and Robson obtained contrast sensitivity functions for many grating types, two of which are shown in Figure 3.4. Note that at intermediate frequencies (changes from black to white), less than 1 percent contrast is needed to resolve a pattern. This is true when sensitivity is measured by varying the contrast of a "square wave" grating (black and white solid stripes, with sharp edges) or by varying the contrasts of a "sine wave" grating (dark and light stripes that fuzz into each other).

INSERT FIGURE 3.4 HERE

To see the effect of contrast on acuity consider the graphs shown in Figure 3.5. The figure shows identical graphs reproduced under different conditions such that one is superimposed on a 'grayish' background, the other on a 'white' background. One can clearly see that it is more difficult to identify the labels on the gray background than it is to identify the labels on the white background. If we assume that black print on a gray background results in a fifty percent contrast reduction relative to black print on a white background, and if the graphs are read in normal reading light (10 ft-L) at a normal reading distance, then equivalent identification accuracy is achieved by increasing the size of the 'black-on-gray' font thirty-one percent (see Figure 3.5). That is, if the size of the label 'vehicle' for the 'black-on-white' font is 0.025 inches, the same label must be 0.033 inches if superimposed on a gray background (using the data in Figure 3.5b).

INSERT FIGURE 3.5 HERE

The effects of contrast are acute in news magazines because their emphasis on "attractive" graphics often results in displayed material appearing on colored or patterned backgrounds. Observing the following rules (Grether and Baker, 1972) will help to increase identification against nonuniform backgrounds. First, choose a color and luminance that contrast most with the colors in the background. For example, a green trend line on a green background will be less discriminable than a red line on a green background. Second, pick light colors for specifiers on dark backgrounds and vice versa. This is necessary because color contrast is not sufficient to ensure legibility; a lightness contrast is far more important (Tinker and Paterson, 1931; Poulton, 1969).

We have mentioned that humans are most sensitive to intermediate degrees of gradualness of luminance gradients across the visual field. As noted, this causes small details to be less resolvable at low contrasts than larger details. However, there is a less obvious corollary of the visual contrast sensitivity function: very gradual changes in lightness will be hard to detect at low contrasts as well. This means that topographic maps and other displays that vary shading continuously across the page may have to use large contrasts if it is desired that the viewer detect gradual changes. For example, in the left panel of Figure 3.6 the change in rainfall across the Great Plains is difficult to detect at the contrast shown; the right panel repairs the problem.

INSERT FIGURE 3.6 HERE

Relative Discrimination

The acquisition of information from charts and graphs often requires one to judge differences in two magnitudes on a single dimension. For example, to acquire information from a bar graph requires that we be able to judge the length of bars. A fundamental question here is how small a difference can be and still be detected. This difference is called the just-noticeable difference, or JND. For bar graphs, then, this means that there will be some minimal difference in the lengths of the bars below which we will be unable to detect differences in length, and thus we will be insensitive to information represented by such differences. Our sensitivity to differences in magnitudes varies from dimension to dimension and is influenced by the context in which the mark is viewed. In the remainder of this section, we present data on sensitivities to various dimensions and also discuss some of the contextual effects.

Difference thresholds are obtained by asking people to compare a test stimulus to a standard and noting how small the difference in magnitudes can be while still being detected. It is a noteworthy fact about human perception that these thresholds depend on the magnitude standard stimulus. For example, using our bar graph illustration, if the length of a bar (the "standard") was 0.1", then very small differences in lengths of another bar (the "comparison") stimuli (say, 0.001") quite possibly would be detected. If, however, the standard were 10.0", then differences of 0.001" would probably never be noticed. Therefore, if difference thresholds are to be useful parameters for guarding against ambiguity, the dependence of threshold and standard must be kept in mind.

E.H. Weber formulated a famous law capturing this dependence in 1846. He showed that the change in stimulus magnitude (ΔS) which was needed to trigger a just-noticeable change in perceived magnitude along any dimension was a constant fraction of the magnitude of stimulation (S) already experienced.[2] Weber's Law means, for example, that if the proportionality constant for bar length was 0.01, then for a standard of 0.1", a comparison stimulus differing from the standard by 0.001" would be detected. For the 10.0" standard, the difference must be 1.0". Table 3.4 lists proportionality constants or differential sensitivities for visual dimensions typically found in graphic displays.

INSERT FIGURE 3.7 AND TABLE 3.4 HERE

The advantage of the Weber fraction as an indicator of differential sensitivity is its independence of the actual units of measurement. For example, it does not matter whether size is measured in inches or centimeters, since both increment ΔS and actual stimulus magnitude S are measured in terms of the same

---

[2] Weber's Law is expressed as:

$$\Delta S = KS \qquad (3)$$

where ΔS is the just noticeable difference (JND). The differential sensitivity to any dimension is obtained from equation (3) by creating the relative quantity $\Delta S/S$, called the Weber Fraction:

$$K = \Delta S/S \qquad (4)$$

Theoretically, when Weber's law is correct, a plot of $\frac{\Delta S}{S}$ versus S results is a constant line as shown in Figure 3.7, with greater ordinate values implying less sensitivity while smaller values indicating greater sensitivity. However, when empirically tested for most sensory modalities, the dashed curve in the figure usually results. At the point S_0, Weber's Law, as written in equation 3, is no longer valid. To cope with this dip in sensitivity occurring near the absolute threshold, alternative laws (Miller, 1947, Guilford, 1932) have been put forth. Miller introduced what now has become known as the generalized Weber's Law.

$$\Delta S = KS + a$$

where the constant a is proportional to the absolute threshold. For most intermediate range stimuli, though, Weber's law holds quite well.

physical quantity, leaving K as a dimensionless ratio. This allows us to compare relative sensitivities for different physical dimensions.

Discrimination of Size

These are numerous different ways of measuring size, each of which will be considered below.

Length discrimination

Length is a commonly used dimension for coding information in graphic displays, especially in coding "point" information as bar graphs do. Our ability to discriminate differences in length is especially important if one must make comparisons across graphs with multiple frameworks. Consider the graph shown in Figure 3.8a, representing yearly fire and police expenditures for some fictitious city. Suppose we are interested in knowing whether fire and police expenditures were the same for the year 1978. Answering this question requires a comparison of the two bar lengths representing these magnitudes. In fact, police expenditures were greater than fire expenditures for that year. However, the difference in the two bar lengths is less than a JND[3], and we can see that it is quite difficult to note the difference reliably (without perhaps the use of a ruler). Figure 3.8b shows the same information, but this time, the difference in length is greater than a JND[4], and it appears much easier to note the difference in lengths.

INSERT FIGURE 3.8 HERE

Ono (1967) investigated the applicability of Weber's Law for line lengths, length being specified both in terms of "physical" size and size of the image

---

[3] If we assume that the bar on the left is the "standard", its length is 1.7". The length of the comparison bar is 1.7625", resulting in a difference in lengths of 0.0625". By Weber's Law and the differential sensitivity to line length of 4.1%, discussed below, the required JND (ΔS) is:

$$\Delta S = 0.041\ (1.7) = 0.0697"$$

[4] The "standard" length is 1.7", and the comparison length is 1.77", the difference being 0.07". This difference is greater than the required JND.

projected onto the retina as measured by degree of visual angle. He found that the value of the JND was predicted equally well for both specifications of size. His results indicate a measure of differential sensitivity to line length of 4.1 percent in terms of either size measure. This means that if one line length is specified at 1", a second line must be specified at 1.041" to be just noticeably different. In terms of the "retinal" size, if one line subtends 1° of visual angle, the second must subtend 1.041° to be perceived as just different.

Orientation Effects. The difficulty of discriminating length is determined, in part, by the orientation of the lines. Consider the graph shown in Figure 3.9a. Suppose a reader is required to make a comparison of the lengths of lines representing the A-C Link (#5) and the G-H Link (#6). Perceptually, the lines appear equal in length. Now let us orient the A-C link in the horizontal position, as shown in Figure 3.9b. We can now clearly see that the G-H link is greater in length than the A-C link. This example demonstrates that differential sensitivity is better for horizontal lines than oblique lines. The same effect is also true for lines oriented vertically. The source of this effect is not optical (Mitchell, Freeman, & Westheimer, 1967), but appears to be somewhere within the neural mechanism involved in spatial resolution (Maffei & Campbell, 1970). We know of no generally applicable quantitative standards concerning the rate of change of differential sensitivity as a function of orientation, but Figure 3.9 does suggest, qualitatively, the direction of this change.

INSERT FIGURE 3.9 HERE

Area Discrimination

Area discrimination is often required for processing information found in "spot" maps. This type of graphic display is often used to represent the number, frequencies, density and the like of variables varying geographically. As an example of area discrimination, consider the "spot" map in Figure 3.10a showing "technology manpower" for different regions in the United States. Manpower is coded in terms of areas of circles: the larger the area, the greater the manpower. Thus, to process the information, we must be able to discriminate between areas. Suppose we wish to compare manpower between the Northeast and Far West regions. If one were actually to measure the diameters of these two circles, one would find that the circle representing manpower in the Far West has a greater area than the circle representing manpower in the Northeast. However, the areas do not differ by a JND[5], making a visual comparison very difficult, if not impossible. The same graph is redrawn in Figure 3.10b with the areas now differing by more than a JND[6], and it is now possible to see the difference.

Baird (1969) has reported a differential sensitivity for area of 6.0 percent. This value implies that for differences in area to be detectable, the areas must differ by 6.0 percent or more.

INSERT FIGURE 3.10 HERE

Discrimination of Number

"Numerosity" refers to the subjective impression of the number of objects that a person can see in the visual field without counting the objects. Our ability to discriminate differences in the number of objects (e.g., dots)

---

[5] The area of the circle representing manpower in the Far West is 0.785 in^2, corresponding to a radius of 0.5". By Weber's Law using the value of differential sensitivity of 0.06, discussed below, the area of the second circle should differ by:

$$\begin{aligned} \Delta S &= 0.06\ (0.785) \\ &= 0.047\ \text{in}^2 \end{aligned}$$

Thus, to be just noticeably smaller, the area of the second circle should be 0.738 in^2 or less, corresponding to a radius of 0.484". The actual radius of this circle is 0.485", corresponding to an area of 0.739 in^2.

[6] The radius of the circle representing manpower in the Far West is still 0.5" (area of 0.785 in^2), but the radius of the second circle is now 0.479 in producing an area of 0.721 in^2, less than the required 0.738 in^2.

becomes important, for example, if one wishes to represent, say, ordinal information concerning population densities of various regions by different dot densities. Taves (1941) established a differential sensitivity index for dot numerosity of 0.204 under nonsimultaneous viewing conditions. This means that if population density of one region is represented by 100 dots, then the density of another region, greater than the first, must be represented by 120 dots to be perceived as just greater. A third region relative to the second should contain 145 dots if it is to be perceived as just different.

Discrimination of Color

Colors may differ in their hue, brightness, and saturation.

Hue

"Hue" is the term referring to the dimension that separates red from green, and so on. Hue is a psychological property, existing in the eye of the beholder. Different hues are produced primarily by differences in the wavelength of light (measured in nanometers, or mμ). Figure 3.11 shows the variation in hue as a function of wavelength along the spectrum, from red through orange, yellow, green, and blue to violet.

INSERT FIGURE 3.11 HERE

Our ability to detect differences in hue is not uniform for equal changes along the physical spectrum. Figure 3.12 shows mean JND's ($\Delta\lambda$) and standard deviations of hue as a function of wavelength (λ) from 410 mμ through 630 mμ (obtained from a set of experiments by Siegel and Dimmick, 1962, and Siegel, 1964). (Recall that the smaller the JND, the greater the sensitivity.) The figure shows that peak sensitivity to hue difference is greatest in the ranges of about 450 to 480 mμ, corresponding to the yellow region. For this region, a change in spectral composition of less than 1.0 mμ is needed to be perceived as "just different". Sensitivity to hue differences is weakest at the extremes of the spectrum (corresponding to the violet and red regions) and also for the green region at about 520 mμ.

We can also see from the figure that the wavelength discrimination function does not at all resemble the function described by Weber's Law (which asserts the JND increases linearly with stimulus level). One possible reason for this is that Weber's law states that the amount of stimulus magnitude that must be added for a JND to be sensed must be proportional to the existing level. That is, discrimination satisfying Weber's Law are mediated by additive perceptual dimensions, such as loudness, whereas color is a substitute dimension. In other words, increasing the wavelength of a patch of light does not lead to the perception of more of something, it leads to the perception of a different something.[7] And, not surprisingly, it has been found that displays that use a gradual shift from one color to another to represent a continuous variable are difficult to understand; we see such variation as a qualitative change rather than as a quantitative gradation (Wainer & Francolini, 1980; Wainer, 1981).

INSERT FIGURE 3.12 HERE

---

[7] These differences mirror differences in the neural substrate of sensation. For additive dimensions like lightness, the magnitude of the stimulus increases, the firing rates of neurons already responding to stimulation increase. It may well be that if this increased firing rate is sufficient, then it results in a JND being experienced. If this phenomenon is not sufficient, then additional neurons are recruited and their added effects evenutally result in a JND.

Hue, on the other hand, is experienced as an attribute of quality in which discrimination is mediated by substitutive processes, that is, which neurons are firing, not simply how many are firing or how frequently they are firing. There are four types of "spectrally opponent" cells responsible for color vision (See Devalois, 1975 for a good discussion.). Briefly, these types are termed red-excitatory, green-inhibitory (+R-G), and yellow-excitatory, blue-inhibitory (+Y-B) and the mirror image of these (+G-R), (+B-Y). Each of these types is spectrally turned to a particular range of wavelength, that is, for certain wavelengths, each responds in an excitatory manner while being inhibited for other wavelengths. As one progresses across the spectrum, there is no additional recruitment of neurons, but instead, a substitution of excitation of one cell for another. Therefore, the sensitivity of whatever cell is firing in response to the stimulus dictates how differences in wavelength are detected. It is interesting to note that the two minima of the wavelength discrimination function occur in the same spectral regions as the "crosspoints" of the two pairs of spectrally opponent cells, the +R-G and +G-R cells having their crosspoints at approximately 590 mμ, and the +B-Y and +Y-B at approximately 500 mμ. If discriminations are based on which cell is firing, then a double minimum in the wavelength discrimination function is exactly what we would expect.

When using color for coding nominal information in graphic displays, Table 3.5 recommends certain hues (coded in the Munsell classification) when fewer than nine colors are needed, which we recommend due to our limited memory capacities. The hues in this table are maximally discriminable from one another.

INSERT TABLE 3.5 HERE

Saturation

If things have the same hue, it is still possible to detect a difference between them because of differences in saturation. Saturation can be thought of as the degree to which a color appears to be rich and pure, free of whiteness, grayness, or blackness. For example, red differs from grayish red in saturation. If light consists entirely of a single wavelength, (say 530mμ, which corresponds to yellow) it is said to be completely homogeneous or monochromatic and has a "colorimetric purity" of 1. White light, on the other hand, is a mixture of all wavelengths, or "maximally heterogeneous" and has purity 0. Between these two extremes exist graduations in purity. If colormetric purity changes with the luminance held constant, the color seems to change principally in grayness. That is, as purity increases, grayness decreases. Colormetric purity, then, is specified as the ratio of monochromatic light in a mixture of monochromatic and achromatic light.

Studies investigating the maximum perceived saturation of various hues (Jones and Lowry, 1926; Priest and Brickweddle, 1938) have shown that saturation appears greatest at the extreme wavelengths and decreases to a minimum at about 570 mμ. Thus, red and blue light will always appear more saturated than yellow light of the same colorimetric purity (i.e., proportion of the light is composed of wavelengths of that hue.)

Experiments by Paneck and Stevens (1966) and Indow and Stevens (1966) have established differential sensitivities for saturation of both primary and intermediate hues. For saturation of red, a primary hue, Paneck and Stevens found that a 2% change in purity is necessary for a just noticeable difference in saturation. Differential sensitivities for changes for hues from 550 to 530 mμ (from a greenish yellow to yellow) and for hues from 630 to 583 mμ (red to yellowish green) were investigated by Indow and Stevens (1966), who found Weber fractions in the range of 2%. Thus, for example, if the purity of color at some dominant wavelength is, say, 80 percent, then purity at the same wavelength must be greater than 81.6 percent, if a noticeable difference in saturation is to be observed.

Because saturation, unlike hue, is perceived as a continuously varying quantity, it is better to use variations of saturation (e.g., between white and richly colored, with pale as an intermediate) then hue in displays like maps where some variable must be plotted as a function of location (Wainer & Francolini, 1980).

Brightness

Colors also differ in their brightness. Brightness discrimination involves the ability to detect changes in luminance along the achromatic scale, black-to-gray-to-white. Lowry (1931) has shown that for maximum discrimination to occur, the luminance of the field should be between 20 and 30 ml. Under this condition, the differential sensitivity is 1.4%. At lower luminance levels, discrimination decreases markedly.

Shape Discrimination

Discrimination of shape is a very complex phenomenon involving sensory, perceptual, and cognitive processes and interactions among these processes. At the higher levels of processing, shape discrimination comes under the headings of form perception or pattern recognition. We will consider here some investigations of our abilities to distinguish changes in relatively simple shapes as certain aspects of the shapes are varied. This may be important for the design of charts and graphs in which the shapes of a set of symbols vary continuously (e.g., from a horizontally-oriented ellipse, through a circle, to a vertically-oriented ellipse) to signal values along some continuum. Similarly, there may be displays in which squares represent one entity, and rectangles represent another entity, so the two shapes must be discriminable if the display is to be unambiguous. In addition, dynamic information such as, for example, an aircraft's glide angle is generally coded as some shape (e.g., a diamond shape) on a cathode ray tube. As the slope changes, the shape changes in its form somewhat. Thus, to maintain a proper glide slope, the ability to recognize changes in shape and discrimination is clearly important.

Veniar (1948) examined subjects' ability to distinguish between a square and a rectangle oriented horizontally or vertically. She established a differential sensitivity of 1.37 percent for shape distortion when either the horizontal or vertical sides of the square were distorted. This value implies that if a 10 cm. X 10 cm. square is projected, a 10.14 cm. X 10 cm. rectangle will be perceived as just different. Veniar also investigated the effects of stimulus area and illumination on discriminability and found no influence of these variables for the ranges considered. Note that this value of differential sensitivity is different from that found for "pure" length discrimination (See Table 3.5), suggesting that different processes may be involved in the two types of discrimination. In fact, in debriefings following the experiment, subjects reported that their judgments involved the shape as a whole, and not the individual line lengths.

In another shape distortion experiment, Kelly and Bliss (1971) investigated sensitivity to distortions of diamond-shaped figures. Distortion was indicated in terms of diamonds appearing "taller" or "shorter" than a standard defined as having a height/width ratio of 1.000. Consequently, diamonds appearing taller and had height/width ratios greater than 1.00 while those appearing smaller had ratios less than 1.00. Kelly and Bliss found a differential sensitivity of 4.8 percent, corresponding to height/width ratios of

1.048/1.000 and 0.952/1.000 for just taller and just shorter diamonds, respectively.

2. The Principle of Perceptual Distortion

Everyone knows that things are not always as they appear. But most people seem to think this is largely due to the occasional optical illusion. However, in many cases there is no illusion but the perceptual system nevertheless is systematically distorting the relationship between the magnitude of the sensation we feel and the value of the physical stimulus property which excited the sensation.

Optical Illusions

Any introductory textbook on perception devotes considerable space to a discussion of illusions (e.g., see Haber and Hershenson, 1982). A number of illusions have been found to affect graph reading per se. For example, Cleveland (1982) found that color on a statistical map can cause an illusion: when colors were highly saturated, a red area was seen as larger than an equal-sized green area; when the colors were not highly saturated, however, no illusion occurred. Another illusion discovered by Cleveland, Diaconis, and McGill (1982) is directly relevant to one of the most common display types: simple scatterplots in which points are plotted within a set of coordinates. Cleveland et. al. asked subjects to judge the degree of "linear association" between the two variables plotted; all subjects had some statistical training and understood the instructions. Judgements were made using a 100 point scale, with 0 being equivalent to r=0 and 100 being equivalent to r=1. When the scale was reduced on the frame, so that the "point-cloud" was reduced in size, subjects saw a higher degree association. This should be evident in Figure 3.12.

INSERT FIGURE 3.12 HERE

An assortment of other illusions may be relevant to special kinds of graphic displays (see Wainer & Thissen, 1981). For example, the "top hat illusion" results in our seeing vertical lines as longer than horizontal lines of the same length. In some exotic plots line-length and orientation can be used to represent information so some lines may be vertical and others horizontal--in which case this illusion would be a source of mis-information. In general, however, most optical illusions usually discussed in perception texts are not likely to affect charts and graphs.

Systematic Distortions: The Power Law

The relationship between the physical magnitude and the psychological magnitude can be expressed by the following formula, due largely to the work of S.S. Stevens.

$$\Psi = k\phi^{b} \tag{4}$$

In the equation above, Ψ is the subjective magnitude of the sensation, Φ is the physical magnitude of the stimulus itself, and b is an exponent (to be determined from empirical data) which characterizes a particular sensory modality (k is simply a constant which relates the units of sensation to those of the physical stimulus property). In other words, for any perceptual continuum, the perceived magnitude of a stimulus is some power function of the stimulus's physical magnitude, with the exact power in the function varying from continuum to continuum. Steven's law is often called the "Power Law" for this reason.

Because the power or exponent in the power function (b) is not necessarily equal to 1.0, sensations often do not change in direct proportion to changes in the physical stimulus. This has some important implications for reading charts and graphs. In the remainder of this section we will discuss the various consequences of the power law for chart and graph comprehension. We shall provide estimates of exponent values for the visual continua commonly employed in producing charts and graphs. Based on these estimates and other research results, we shall make recommendations on how best to use these continua in visual displays. In many instances, there is considerable variation in estimates of exponents, due to differences in research methodology, in which cases we provide the range of values.

General Consequence of the Power Law

The form of equation 4 has several mathematical properties reflecting properties of perceptual systems that may have enhanced the species's chances for survival in a natural environment. First, the power law provides for ratio invariance. That is, equal stimulus ratios induce equal sensation ratios As a consequence, an object in the environment appears to retain a constant size and shape in relation to background objects as its position changes relative to the observer. Second, for some sense modalities, such as visual brightness, the natural environment may present a broad range of values (up to 10 orders of magnitude). If the visual system were to transduce and process brightness information linearly (i.e., the exponent were 1), the system would have to be much larger to register the entire continuum and probably would have to possess a greater neuronal mass. In fact, however, the psychophysical exponent for brightness is less than one. This enables the same range of physical brightness to be registered within a smaller sensory system. The information that is lost by virtue of the nonlinear sensory transformation has little importance for survival.

These two advantages provide an explanation of why the evolutionary process has favored a power function for sensory encoding. In terms of graphical applications, however, the consequences may not be so happy: the power law can distort the presentation of information when the continuum that is being used has an exponent greater than or less than 1.

The perceptual distortion that occurs as a result of the psychophysical power law can best be explained using a graph. Figure 3.13a shows a power function with an exponent of .7 (i.e., the physical magnitude is raised to that power to predict the corresponding subjective magnitude). As a consequence of the exponent being less than one the value of the sensation, ψ, increases less rapidly than that of the stimulus, ϕ. Figure 3.13b shows a graph of a power function where the exponent, b, is 1.2. In this case, the value of the sensation, ψ, increases more rapidly than that of the stimulus property. Finally, Figure 3.13c shows a plot of the power function whose exponent is 1.0. This function is represented as a straight line. In this and only this case no perceptual distortion occurs, with sensation increasing at the same rate as the stimulus magnitude. Thus, only when the exponent is equal to 1.0 do things differ in the way they appear to differ.

INSERT FIGURE 3.13 HERE

The foregoing characteristics of the sensory power law have important implications for graph construction and comprehension. First, equal intervals on a stimulus continuum do not, in general, correspond to equal intervals on the subjective continuum. This is made clear in Figures 3.13a and 3.13b where w_1 and w_2 are equal intervals on the physical continuum ϕ. The corresponding subjective intervals W_1 and W_2 are obviously not equal ($W_1 > W_2$). Equal physical intervals correspond to equal subjective intervals only for sensory dimensions whose power functions have an exponent of 1.0. In the construction of graphs, one is often concerned with conveying a relation between two quantities by using a corresponding visual relation between graphic symbols representing those qualities. When the information to be conveyed is interval scaled, the graph maker should be aware that equal intervals on the scale of interest may not portray equal subjective intervals on the particular visual continuum.

For example, suppose we wish to construct a chart that provides information on various occupations. Each occupation is to be represented by a circular area of uniform size and the mean intelligence quotient (I.Q.) for people in each occupation is to be indicated by the apparent lightness of a particular red hue used to color each circle. Consider three occupations, A, B, and C, with mean I.Q.'s of 130, 110, and 90, respectively, which are represented with a 622.nm red color with a colorimetric purity of .51. Table 3.6 provides a list of the exponents for perceived lightness as a function of the corresponding physical stimulus property, reflectance, considered separately for various colorimetric purities for different hues. Assuming that the red graph is to be read under artificial light, an exponent of .62 is appropriate. Using an arbitrary constant (k=1.0) for the formula (in equation 4), lightness is plotted against reflectance in Figure 3.14.

INSERT TABLE 3.6 AND FIGURE 3.14 HERE

Because the three occupations are spaced at equal (20 point) intervals we must choose our reflectance such that the intervals in subjective lightness are also equal. The symbols on the chart can be made most discriminable if the full range of available reflectances are used. Similarly, the meaning of the chart is made most transparent if lighter shades of red are assigned to occupation groups with the brighter people (for reasons to be discussed later). Thus, we select a value of .3 on the subjective lightness scale for occupation A, a value of .2 for occupation B, and .1 for occupation C. The positions marked I_A, I_B, and I_C on Figure 3.14 indicate the corresponding points on the psychophysical function. From this it can be seen that reflectance values of .170, .0[illegible], and .034 must be employed for occupations A, B, and C, respectively.

Although this set of values satisfies the requirement that equal intervals in the referent scale (I.Q.) are represented as equal intervals on the subjective lightness scale, it is certainly not unique in this respect. We might have selected a subjective lightness value of .29 for occupation A and used .9 unit decrements on the lightness scale for the other occupations. The graph maker has considerable latitude in following this procedure to remove distortion in interval scaled graphs. In addition, the level of precision discussed here may exceed that needed for most uses of most charts and graphs. But in all cases, the graph maker should be aware that equal differences in physical units may not be seen as equal and hence may not function to communicate effectively.

Of course, in some cases, adjusting the elements of the graph to compensate for the distortions of the human perceptual system may be the wrong thing to do. For example, for elements such as squares with different areas, the reader may want actually to measure the elements to obtain absolute value, to interpolate, or to verify that the graph really supports the claims made in the accompanying text. In such a case, altering the areas so that they differ from the exact quantitative values dictated by the information being communicated would have disastrous results. Note that simply by choosing a continuum with an exponent of 1.0 to begin with to represent the quantities, the graph maker thereby avoids both perceptual and physical distortions.

The second major implication of the power law for chart and graph construction concerns the property of ratio invariance, which was mentioned earlier. Recall that "ratio invariance" means that if the ratio between the values of two stimuli is equal to the ratio between the values of two other stimuli, then the ratios of the corresponding pairs of sensations are also equal.[8] A

---

[8] This can be demonstrated quite succinctly by the following equations. Let ϕ_1, ϕ_2, ϕ_3 and ϕ_4 represent values on a physical continuum for four stimuli, and let ψ_1, ψ_2, ψ_3 and ψ_4 represent the sensations which result from these stimulus values. Then if:

$$\frac{\phi_1}{\phi_2} = \frac{\phi_3}{\phi_4} \qquad \text{(a)}$$

it follows that:

$$\frac{\psi_1}{\psi_2} = \frac{\psi_3}{\psi_4} \qquad \text{(b)}$$

This is most easily seen by employing the power law, $\psi = k\phi^b$, to rewrite equation (b) in terms of equation stimulus properties.

$$\frac{k_1\phi_1^{\,b}}{k\phi_2^{\,b}} = \frac{k\phi_3^{\,b}}{k\phi_4^{\,b}}$$

Raising boths sides of equation (c) to the (1/b)th power yields equation (a), as promised.

$$\frac{\phi_1}{\phi_2} = \frac{\phi_3}{\phi_4} \qquad \text{(d)}$$

particular ratio of stimulus properties will not, in general, yield the same ratio of sensation magnitudes. Given a stimulus ratio, (ϕ_1/ϕ_2), the resulting sensation ratio, (ψ_1/ψ_2), is $(\phi_1/\phi_2)^b$. Thus, ratios of stimulus properties are also transformed by the power law. A given stimulus ratio can produce the same sensation ratio only when the sensory system involved is characterized by an exponent of 1.0.

Let us suppose, then, that we wish to convey the idea that country A has x times the population of country B by sketching outline maps of the two countries, where the apparent sizes of the drawings reflect the corresponding population sizes. In order to accomplish this graphically, we would employ the exponent b in the psychophysical relationship between the physical property, area, and its subjective correlate, apparent size, as follows:

$$x = \frac{\phi_A^b}{\phi_B^b} \qquad \text{or} \qquad x^{1/b} = \frac{\phi_A}{\phi_B} \tag{5}$$

If we select an appropriate area, ϕ_A, for country A, then equation (5) can be used to determine ϕ_B such that the ratio of apparent sizes is x.

$$\phi_B = \phi_A/(x)^{1/b} \tag{6}$$

The non-linearity of the psychophysical power law, therefore, has implications on the use of graphic symbols to portray relationships between symbol referents on intervals and ratio scales. Knowledge of the values of the power law exponents for the various visual continua, together with an understanding of their role in perception, is important for accurate graphic communication.

The power law does not only bring bad news to the graph maker, however. Simple (i.e., nonpolynomial) power functions are monotonic: they either increase or decrease along their domains, but never both. That means that ordinal relationships (e.g., A is larger than B, but not by any particular amount or ratio) will virtually always be perceived veridically if conveyed by some physical continuum. If all the graph maker wishes to convey is the ordinal relationships among a set of entities, and not their exact differences or ratios, the power law does not entail any problems.

A few words of caution are in order before proceeding with a discussion of the systematic distortions to be found in the visual continua that are important to chart and graph construction. In many cases there is considerable variation in the exponents that were determined empirically for a given continuum by different investigators. At first glance this variation would appear to be random. However, the results obtained from a given type of measurement procedure are sometimes known to be consistently higher or lower than those from another procedure. Because much of the variation in the tabulated exponents is, in this sense, systematic, the reader should view the results to follow with reduced skepticism.

Most of the estimates we provide were obtained either with an "estimation procedure" or with a "production procedure". In the estimation procedures, a subject is first shown a standard stimulus to which the experimenter assigns a number, and the subject is then asked to respond to each of a series of experimental stimuli by saying a number which reflects the subjective or perceived relationship of the experimental stimulus to the standard stimulus. In the production procedures, the experimenter presents the standard, and gives the subject the means to vary stimuli (e.g., by turning a knob controlling the brightness of a light) to yield a given relation to the standard (e.g., five times as bright). In general, estimation procedures yield lower exponents than production procedures. The availability of the standard stimulus and its magnitude in relation to the experimental stimuli also have systematic effects on the exponent later computed.

With this in mind, let us now specifically examine the visual continua which are most often used in graphic representation.

Line Length and Inclination

The relationship between physical line length and its subjective correlate, apparent line length, has been studied thoroughly by a number of investigators. Table 3.7 presents the results of some of these studies, as well as some of experimental conditions and methods employed. In all cases the exponent was found to be close to 1.0, thus indicating a linear relationship between the physical and subjective continua. Although the veridicality of length perception is not surprising, it is, nonetheless, a fortuitous result because it obviates the difficulties of employing nonlinear transformations in order to encode information so that it is perceived accurately.

INSERT TABLE 3.7 HERE

In light of this finding it is easy to understand why bar graphs are so pervasive as a graphic format. Information in a bar graph is encoded directly by the lengths of a set of discrete lines or bars. Because perception of ratios is veridical in this case, the information can be assimilated without resort to any mental or graphic gymnastics.

Stevens and Galanter (1957) found that the relation between angular orientation of a line and subjective inclination is also linear. And as before, Miller and Sheldon (1969) obtained a linear relation between the average inclination of a group of six lines of varying orientations and the subjective average as perceived by their subjects.

Thus, we may conclude that straight lines are well behaved in a psychological sense. The relation between physical attributes, such as length and orientation, and the subjective correlates of these attributes is linear, even when average quantities describing groups of lines are at issue.

Area

Unfortunately, the simple relation that obtains for the actual and perceived length of lines is lost when lines enclose areas. The relationship between physical area and apparent size of various two dimensional figures has been studied extensively by numerous investigators. The results of some of these studies, organized by the shape of the figure, are presented in Table 3.8.

INSERT TABLE 3.8 HERE

Initial perusal of this table leaves one befuddled by the wide range of exponent estimates (from .55 to 1.20). However, much of this variation can be attributed to identifiable sources, many of which have implications for graph construction. One important influence on the exponent value is the instructions given to subjects prior to the experimental task. Teghtsoonian (1965) asked half of her subjects to "estimate the apparent sizes" of a set of circles. The other half were asked to "base their judgements on the actual physical areas". The exponent resulting from apparent size instructions was .76, whereas that resulting from physical size instructions was 1.03 (see also Macmillan et. al., (1974); Teghtsoonian, 1965).

These results suggest some guidelines for the graph maker. Under certain circumstances a person's perception of the ratio of two physical areas is nearly linearly related to the actual ratio of physical areas. These circumstances are: 1) the person is specifically asked to attend to physical area; 2) the person understands the concept of area and how it is calculated; 3) the stimuli are similar in shape and the shapes possess enough linear cues to enable an accurate area estimate. Under these conditions, the exponent ranges from .75 to 1.0. In contrast, if a person is instructed to (or will spontaneously) attend to apparent size, or the shapes are nonsimilar and irregular, or the subject is mathematically naive, exponents may be substantially lower. A person who interprets apparent size to be indicated by a prominent linear dimension will operate with an exponent near .5; however, most people will operate with exponents between .7 and .8. -- resulting in reader's systematically under-estimating differences among increasing areas.

Thus, given the incompleteness of present theories of how we compare the areas of differently-shaped figures, the graph maker is advised to avoid using different shaped figures when precise interval area relationships are to be communicated. Sadly, this could apply to the currently popular maps in which the magnitude of some attribute of a country (e.g., oil reserves) is conveyed by the size of the country on the map.

Volume as Implied by Perspective Drawings and Volume in Real Space

Perspective drawings of solid objects are frequently used as symbols in charts and graphs. The use of these drawings has been of particular interest to cartographers who use symbols to convey simultaneously information about a location and some other attribute, such as population, of cities on maps. For them, the use of simple area symbols for cities becomes unwieldy because heavily populated cities require inordinately large areas which would imprecisely mark the location of the city and possibly obscure smaller cities in the vicinity. Implied volume is one way of overcoming the problem. For example, perspective drawings of two cubes, one with 1 mm edges and one with 10 mm edges, imply a volume ratio of 1000:1. On the other hand, simple squares with 1 mm and 10 mm sides imply an area ratio of only 100:1. The availability of this theoretical advantage is of course contingent on readers having the ability to estimate ratios of volumes in real space accurately, and on their perceiving a perspective drawing of a solid as they would an actual solid. Therefore, the exponent for volume in real space should be close to 1 (at least when subjects are asked to attend to actual physical volume rather than apparent size) and the exponents for actual solids and perspective drawings of solids should be nearly equal.

Exponents for the psychophysical relation between physical and subjective volume for real solids have been estimated by Ekman and Junge (1961) and by Teghtsoonian (1965). These are presented in Table 3.9. Exponents for the relation between physical and subjective volume as implied by perspective drawings have been estimated by Ekman and Junge (1960), (1961) and by Ekman, Lindman, and William-Olson (1961). These are presented in Table 3.10.

INSERT TABLES 3.9 AND 3.10 HERE

The conclusion to be drawn from experiments on drawings of cubes (obviously of more interest to the graph maker than real cubes), then, is that most people compare small perspective drawings of three dimensional objects on the basis of the area enclosed by the drawing and not by the actual volume implied. The graph maker, therefore, should not attempt to employ perspective drawings with the expectation that readers will perceive differences in volume veridically.

Proportion and Numerousity

The concept of proportion is often conveyed graphically by a pie chart. Radii at various inclinations divide a circle into segments, and proportion information is encoded primarily by the relative areas of the segments. Because the perceived inclination of a line is linear in relation to the actual inclination, as previously discussed, the pie chart should be effective in conveying proportion information. The display format of the pie chart is very rigidly structured, however, and not conducive to conveying information in addition to proportions. If, for instance, the chart is to be used to show the proportions of different ethnic groups in the U.S., all members of each group must be gathered up, in a sense, and put in the appropriate segment. Information about the proportion and the geographic distribution of group members cannot be conveyed simultaneously.

Studies have been reported by Stevens and Galanter (1957) and by Rule (1968) on a subject's ability to estimate proportions in a less structured format. Although these investigators were not specifically concerned with graph and chart comprehension, their results can certainly be applied to this issue. In the Stevens and Galanter study the stimuli were blue and green dots placed randomly in an 8 cm square. The total number of dots was 36 in all cases but the proportion of blue to green was altered for the various stimuli. When subjects were asked to estimate the percentage of blue and green dots, estimates were most accurate at the two ends and at the center of the stimulus range. However, the subjects' percentage estimates increased linearly with increases in the actual proportions. In a similar experiment, Rule asked subjects for magnitude estimates of the proportion of dots and lines occupying the positions of an eight by ten rectangular array and obtained an exponent of .97.

Numerousity, as noted earlier, refers to one's subjective impression of the number of elements in some collection gauged without counting these elements one-by-one. This continuum differs from that of proportion in that numerousity is concerned with elements of one type, the number of which is neither confined to a specific range nor considered in relation to the number of some other type of elements. Estimates of the exponent for the perception of numerousity range from .65 (Taves, 1941) to 1.34 (Stevens, 1957), reflecting differences in methodologies. Krueger (1972), examining the perception of numerousity and how it is affected by display size, offered an exponent of .85 as his best estimate for a true exponent for numerousity, averaging across estimation (.72-.78) and production (.93) methods. The exponent of .85 for numerosity indicates that subjects typically under-estimate the number of items present. Items displayed in a compact area tend to be underestimated more than those shown in a large format (owing mainly to a difference in the proportionality constant, K (e.g., 4), not the exponent), but the effect of the size of the format seems to diminish (or saturate) at some point. The exponents for proportion and numerousity are presented in Table 3.11.

INSERT TABLE 3.11 HERE

Lightness and Saturation

In many instances, differences in color attributes or lightness of gray tones serve as a basis for differentiating chart and graph symbols. We, therefore, include a brief discussion of the power law exponents of various chromatic and achromatic attributes of visual stimuli.

In order to minimize ambiguity in this presentation, recall our earlier use of some relevant terms. Hue is the attribute of a color perception denoted by the names blue, green, yellow, red, purple, etc. An achromatic color perception is one which possesses no hue (e.g., white, gray, and black). Saturation is the attribute of color perception determining the degree of difference from the achromatic color most resembling it. Brightness (of an area perceived as self luminous, such as a computer video display) is the perceptual dimension ranging from very dim to very bright or dazzling. Lightness (of an object perceived as non-self luminous, such as a price of paper) is the perceptual dimension ranging from dark (black, for achromatic stimuli), to light (white, for achromatic stimuli). Recall that each of these subjective continua is associated primarily with a physical continuum. Hue is chiefly associated with wavelength, saturation with colorimetric purity, brightness with luminance, and lightness with the luminance factor (percent of incident light of what the surface reflects back). More complex relations are also operative in color vision; for instance, hue is affected somewhat by purity and luminance, and brightness is affected somewhat by wavelength and purity.

Guirao and de Mattiello (1974), using non-self luminous surfaces, obtained exponents reproduced in Table 3.12. Note that the exponents for small-sized fields are greater than those for the large fields, regardless of the type of illumination, for all hues except yellow. Also, blue, green and red have lower exponents when viewed under daylight conditions than under artificial light.

Yellow and orange appear to be unaffected by the type of illumination. Self luminous colors, such as those which appear on a computer graphics screen, are characterized by lower exponents than the surface colors (see Indow & Stevens, 1966), although these exponents generally are over 1.0.

INSERT TABLE 3.12 HERE

In another study, de Mattiello and Guirao (1974) examined the relation between lightness, luminance factor (% reflectance), and colorimetric purity. The exponents they obtained for lightness as a power function of percent reflectance at a given colorimetric purity are presented in Table 3.13.

Although no studies have been performed to determine whether saturation exponents change continuously with the size of the color patch, the graph maker should be aware that saturation may be affected by the size of a colored figure, with greater exponents for smaller areas. It is almost as if the same color placed in a smaller area appears "denser" and hence, more saturated. Thus, slight differences in colorimetric purity may be required to make two figures of the same hue but different sizes appear equal in saturation.

INSERT TABLE 3.13 HERE

Conclusion

In closing this section, it is worth noting that the foregoing principles allow us to explain some data collected on graph reading per se. Croxton and Stein (1932) examined the ability of bar graphs to convey the relative magnitude of two quantities. They compared the accuracy of subject's estimates of the ratio of: (1) one bar length to another, (2) one square area to another, (3) one circle area to another, and (4) one cube volume (as depicted in a line drawing) to another. They found that accuracy of the estimates decreased with increasing number of dimensions; bars were more accurately compared than squares, circles or cubes; squares and circles were more accurately compared than cubes; and squares and circles were equally well compared. These results are not surprising. Given the fact line length is known to be a linear function of physical line length but perceived area and perceived volume are nonlinear power functions of physical area and volume, respectively, with the exponent of the volume power function deviating further from 1.0 than that of the area power function. Croxton and Stein also compared the accuracy of estimates of the relative arcs of pairs of circles and squares when the centers of the paired figures were horizontally aligned versus when the bases of the figures were horizontally aligned. No difference was found, as we would expect from the foregoing discussion (see MacDonald-Ross, 1977, for an extensive review of this literature).

Thus, the material reviewed in the preceding sections chapter can serve as a substitute for a vast number of potential experiments on the accuracy of reading graphs with different sorts of physical marks. Where the Weber fraction for a sensory continuum is small and the exponent is close to 1.0, we can expect more accurate interpretation than we would for other continua. Even better, we can predict the types of errors that will be made, the sorts of adjustments that can eliminate the errors and how the type and extent of distortion varies with extraneous factors such as size and illumination.

II. Natural Units

1. Gestalt Principles

If a display is to be read accurately, the marks must be read and organized correctly. So far we have been concerned with factors that must be considered if the marks themselves are to be read correctly, and now we turn to factors that underlie how we organize marks into psychological units. As we shall see when we turn to the graph comprehension model outlined in Chapter 6, how the visual system parses the visual input into units and links these units together has important consequences for how easily the various parts of the graph are recognized and how easily the appropriate quantitative and conceptual information will be extracted from it. This section will review briefly the progress made in the study of perceptual organization and consider the implications for graph construction.

The Gestalt psychologists began work in the 1930's that has led to some genuine understanding of how visual stimuli are organized perceptually. The Gestalt psychologists believed that visual stimulation initiates the action of organizing electrical forces in the nervous system, which separate figures from their backgrounds, establish distinct groups of objects, and define structure in the visual scene. Although many of the physiological models postulated by the Gestalt psychologists have since been proved incorrect and many of their explanations of organizational processes have been found inadequate, some key features of their approach to visual organization continue to be of value in the study of perception. Many researchers now believe that the Gestalt Principles reflect the operation of mechanisms that seek to carve the continuous optic array into distinct portions, each of which corresponds to a physical object in the world. Although charts and graphs are not typical visual stimuli found in the world, we can expect that these same mechanisms will go to work on graphic stimuli and attempt to discern "objects" and their interrelationships in them.

Max Wertheimer (1938) formulated a set of "laws" (or principles) of organization of visual forms involving the following factors: proximity, similarity, continuity, closedness, and symmetry. Several of these factors may be operative in the same scene. In some cases, all factors may be cooperating to emphasize a common structure. In other cases factors may be set in opposition to each other, with each promoting a different structure. On such occasions, one of the alternate structures usually emerges as dominant, although weakened by the conflict. The major shortcoming of Wertheimer's principles for our purposes becomes apparent here. That is, when two factors are in conflict, Gestalt Theory cannot predict which will emerge victorious. In the ensuing discussions of each organizational factor, we shall provide examples of cooperation and conflict to illustrate the operation of these principles in charts and graphs.

Proximity

Figures that are situated near each other tend to be associated with each other to form a common structure. This is demonstrated quite clearly in the examples shown in Figures 3.14a and b. The spacing between the circles of Figure 3.14a induces the observer to group pairs of circles together in the pattern 12/34/56/etc. In Figure 3.14b the observer is lead to group the squares into triplets: 123/456/789/etc. In these simple examples the influence of proximity is so strong as to preclude alternative groupings such as 1/23/45/67 in Figure 3.14a.

INSERT FIGURES 3.14 AND 3.15 HERE

Grouping by proximity is easily studied because it is one of the few Gestalt principles where it is possible to obtain quantitative measures of the stimulus properties. For example, Kohler and Adams (1961) used an array similar to that shown in Figure 3.15, but varied the ratio of spacing between rows and columns (from 1.0, equal distances, to .25, where inter-row space is 4 times inter-column space). If subjects were not directly attending to the display, the stimulus ratio necessary to induce them to report row or column organization was about .38. If subjects were directly attending to the display, however, and they were looking for row or column organization, a ratio of about .62 was required for articulation. In other words, when a person actively looks for a particular sort of organization, the stimuli need not be physically separated to as great an extent as when a person has no prior organization in mind when first seeing a display. These figures then, give some rough estimates of how to space a field of patterns to use proximity to engender an organization into rows and columns, depending on whether the reader is expected to anticipate a given organization in a graph or not.

Proximity is one of the factors most commonly used to organize a chart or graph. For instance, except when a remote key or legend is employed, proximity is the usual means by which labels are associated with their referents. Figure 3.16 shows an example of a graph extracted from one of the national news magazines in which the labels for the vertical axes are located at some distance from the axes to which they correspond. In this case, confusion results, although it is not severe and can be resolved in a short time. This failure to use proximity to express association, however innocent, exacts a cost from the reader in his attempt to understand the graph.

INSERT FIGURES 3.16 AND 3.17 HERE

A more severe problem can ensue when proximity is misused so that an unintended structure emerges. Consider, for example, Figure 3.17a. In this case a 5 by 5 array of circles is employed to show the fraction (column labels) of the average daily requirement of various nutrients (row labels) supplied by a serving of a given food substance. Because the distinct entities to be scaled are the various nutrients (one nutrient per row), and not the various proportions in their own right (which have no intrinsic interest), the semantics of the chart requires an organization into rows. The relative spacing between circles in the vertical and horizontal directions, however, clearly induces a perceptual organization into columns. This renders the information in the graph obscure until the conflict between the semantics and syntax is resolved. Figure 3.17b shows this chart redrawn with the proximity of circles favoring articulation by rows. Note that the meaning is much more evident when proximity is properly employed.

INSERT FIGURES 3.18, 3.19, AND 3.20 HERE

Similarity

Figures which resemble each other tend to be seen as grouped together. Figure 3.18a and b show an example of similarity acting as a grouping factor. The 12/34/56/78 pattern emerges clearly in 3.18a as does the 123/456/789/10 pattern in 3.18b. Note that in 3.18b this structure emerges in spite of a stimulus spacing which favors a 12/34/56 structure. For estimates of the strength of grouping by various sorts of similarity (brightness, shape, hue, etc.) relative to proximity grouping, see Hochberg and Silverstein (1956) and Hochberg and Hardy (1960), some of whose results are summarized in Figure 3.19.

Similarity can be quite useful in expressing a relationship between graphic elements which, because of the format of the graph, must be situated at some distance from each other. For instance, Figure 3.20 shows a series of three graphs, each of which contains three specifier elements. In this case, although the 3 graphs discuss different topics (as indicated by their titles), each topic is discussed in common terms: a normal range, an actual level, and a minimum acceptable level. Although the specifier elements in each case are appropriately and adequately labeled, the similarity in form and texture of the elements serving the same function in each graph visually emphasizes the common semantic interpretation. In fact, once the specifier elements of one graph have been identified and interpreted, the meaning of the elements in the two remaining graphs becomes immediately obvious. The alphanumeric labels of the specifier elements in these remaining graphs assume only secondary importance in the presence of the similarity grouping factor.

As with proximity, similarity can be misapplied to suggest an unintended structure. Consider the modification of the previous series of graphs shown in Figure 3.21. In this series the alphanumeric label "ACTUAL LEVEL" is printed in large bold face type supposedly to emphasize the importance of this element relative to the other two. At first glance, however, this type face closely resembles that of the graph titles, "CRUDE-OIL STOCKS", etc. Only upon close scrutiny can the reader discern the slight differences in aspect and slant. The net result of this unintended similarity is that the reader may at first be led to believe that this label is part of the graph title and the reader pays a price, albeit small, in time and effort to correct this false implication.

INSERT FIGURES 3.21, 3.22, AND 3.23 HERE

Although Figure 3.19a shows similarity and proximity competing to promote different structures, these factors can be made to cooperate in emphasizing a single structure. Figure 3.22 shows such a situation.

In general, it has been found in studies of tabular and textual materials that the use of similarity and proximity in providing redundant information about hierarchical organization benefits readability. For example, labels indicating subdivisions at the same hierarchical level should be similar in their left-to-right placement on the page, size of type, boldness of type, and case (Wright, in press). The subdivisions themselves in a list are best when set off so that their left margins are aligned, so that the pieces in a subdivision cohere because of proximity, common fate, and good form (see below) (Hartley, 1978; Stewart, 1976; Waller, 1977). For a graph maker similar patterns of readability can be expected: subgraphs or groups of lines or bars that are related at a particular level in a conceptual hierarchy (e.g., differ ent years, different seasons within a year, different months within a season) should be near each other, similar to each other, aligned with each other, and should bear alphanumeric labels with similar visual characteristics. Conversely parts of graphs that belong to different groupings should differ along these dimensions.

Although the redundant application of grouping factors can be a powerful tool in graphic representation, its misapplication can severely obscure the meaning of a graph or chart. Figure 3.23 shows a chart similar to that of Figure 3.17 but with a particularly unfortunate coincidental distribution of nutrients. In this case the proximity and similarity factors cooperate to render any but columnwise organization inaccessible.

INSERT FIGURE 3.24 HERE

Symmetry

A symmetric arrangement of marks is more likely to be interpreted as a figure than the same marks in the absence of the symmetric relationship. The operation of symmetry as an organizing force is demonstrated in Figure 3.24. The parts labelled a, b, and c in the left half of the illustration stand out as white figure on a dark background whereas the corresponding forms, d, e, and f in the right half of the illustration appear as dark figures on a light background.

As for possible applications of symmetry grouping to charts and graphs, it is noteworthy that a symmetric arrangement is the conventional format for presenting a key, legend or table. The overall symmetry of these items first identifies them as simple structures, and secondly establishes the desired correspondence between opposing elements. Compare the keys on both panels of Figure 3.25; the left is clearly superior.

Figures 3.26a and b demonstrate another use of symmetry; that of grouping together different graph elements to emphasize the convergence or divergence of the pair, rather than the individual trends of each element. In these figures the vertical arrangement of the two subgraphs in combination with the unconventional location of the horizontal axis in the upper subgraphs creates a symmetry which draws attention to the higher-order relationship between the specifier elements. Figures 3.27a and b show alternative representations of the information contained in the subgraphs of 3.26b. Note that the effect of these presentations is much less striking.

INSERT FIGURES 3.25, 3.26, AND 3.27 HERE

Good Continuation

When presented with a configuration of discrete marks or a set of curved lines that cross each other or double back on themselves, an observer will perceive the organization in which the elements are as smooth and continuous as possible. Figure 3.28 a, b, c, d, e and f show examples of configurations in which this phenomenon operates. Note, for example that the discrete points in 3.28a appear to be structured as distinct straight or smoothly curved continuous line segments. Perhaps the most likely organization to be perceived in 3.28a is that shown by solid lines in 3.28b. Again in 3.28c one is most likely to perceive a 12/3 structure rather than the possible alternatives (13/2 or 23/1) because this dominant structure avoids sharp changes in line direction. Similarly, the organization 13/24 emerges in 3.28d for the same reason. Figures 3.28e and f show cases in which the factor of good continuation is made to compete with closedness (to be discussed later). Good continuation dominates in both figures to create the structure 1 3 5 7.../2 4 6 8...

INSERT FIGURES 3.28 AND 3.29 HERE

The graph maker can exploit good continuation to link a label with its associated specifier element. Figure 3.29 shows an extreme case, where grouping by good continuation can actually overcome grouping by proximity to associate labels with their corresponding lines. Furthermore, good continuation can tell the graph maker when it will be necessary to differentiate two lines on the basis of color, dots vs. dashes, etc.

When the segments of the lines can only be grouped in one way that conforms to the principle of good continuation (e.g. Figure 3.30a), differentiating the lines does not add appreciably to the readability of the graph. However, when the lines have similar slopes where they intersect, good continuation will not favor one organization over another, and ambiguity will result unless the appropriate line segments are linked to one another by the principle of similarity (compare Figure 3.30b to Figure 3.30c). We have found that a recurring cause of ambiguity in line graphs is the perceptual mis-segmentation of close, intersecting lines. The law of grouping by good continuation explains this ambiguity and should alert the graph maker to this potential pitfall.

INSERT FIGURE 3.30 HERE

Common Fate

According to the Law of Common Fate, elements in a moving display that are moving in the same direction and at the same velocity will be grouped together. In a stationary display, lines that follow the same trajectory across the page (i.e., are parallel to one another) will be grouped together. Thus the curved lines A, B, C, and D in Figure 3.31 will be grouped together despite being dissimilar and far from each other. Like good continuation, common fate can help the graph reader to link labels to their associated lines, even if a label cannot be placed at the end of its line or closer to it than to other lines.

INSERT FIGURES 3.31 AND 3.32 HERE

In relational graphs displaying discrete data, such as is shown in Figure 3.32, a graph maker is sometimes tempted to emphasize a possible trend or relationship by the addition of a continuous line through the swarm of points. The principle of grouping by common fate makes it likely that the line and the subset or envelope of points that follow the same curve as the line will be grouped perceptually. Grouping by common fate might also allow the possibility of abuses in certain cases, given that the visual system may tend to group the dots into a structure with the same trend as the line even when no such trend exists in the actual geometry of the points themselves. The honesty of such an addition then depends on the particular situation. The graphmaker, in general, owes the graph reader some justification for superimposing a trend line onto a collection of dots. If the continuous line represents a locus of prediction based on some theory or a regression line resulting from statistical analysis, this should be clearly stated on the graph or in the associated text.

INSERT FIGURE 3.33 HERE

In other instances, a graph may display a sequence of discrete data such as a time series. Such graphs are often drawn with sequential points connected by short intervening straight lines, Figure 3.33 shows a time series with and without connecting lines. Although this connected dot display format has the advantage of making the sequence of dots easier to follow, Rouse (1974) has shown that such a format makes it more difficult to estimate visually the standard deviation about the mean of the data. Subjects systematically underestimated this statistic for both the connected and unconnected dot formats, but their estimates using dots connected by straight lines were significantly worse. Most likely, by grouping the dots into a continuous curve, common fate simultaneously made the dots harder to see as a swarm of dots per se (because Gestalt organizations compete with each other rather than coexisting perceptually). And if it is hard to see the points as a swarm, it will be hard to see the properties of the swarm (e.g., its average width) that are necessary to estimate its standard deviation.

INSERT FIGURE 3.34 HERE

Good Figure

A region that is defined by a closed boundary is more likely to be seen as a figure than one which is incompletely closed or left open, especially if the resulting shape is simple and regular. This can be seen in the illustrations of Figure 3.34. The progression of drawings from left to right in Figure 3.34a becomes more figure-like as the boundary approaches closure. The effect of closure on natural organization is made evident in Figure 3.34b where the dominant structure emerges as 12/34/56. A portion of this figure, left open by the absence of elements 5 and 6, reverts to a 23/14 structure under the influence of good continuation. Finally, Figure 3.34c shows a modification of a previous figure where the interruption of good continuation allows closure to dominate the figural organization.

Good form is employed as an organizing force in charts and graphs in a variety of ways. A closed boundary may be used to separate a graph or series of graphs from surrounding material such as text. The outer framework of a graph may completely or partially enclose the specifier and, thereby, define the coordinate system with respect to which the specifier elements of the graph are described (see Chapter 6). The complete closure provided by the circular framework of a pie chart serves this function, as does the partial closure implied by the orthogonal axes of a Cartesian framework. It is not unusual for closure to be completed in a Cartesian system by a background field of homogenous hue (Figure 3.35a) or by two thin lines completing the rectangle defined by the principal axes (Figure 3.35b). Closedness is employed within a graph to define sub-structures such as keys and legends (Figure 3.36) and small specifier elements appear more substantial and noticeable when defined by closed contours. Also, closed symbols and characters are more dominant than open ones. Observe that the letter "o" appears more dominant than the letter "c" in Figure 3.37.

INSERT FIGURES 3.35, 3.36 AND 3.37 HERE

As with the preceding organizing factors, closedness can be and sometimes is used to disadvantage in graphs. Figure 3.38 shows a graph which was extracted from a leading national news magazine. The point of this graph is to show that wholesale prices (upper subgraph) are increasing while the value of the dollar relative to the German mark is dropping. The divergence of the jagged lines lines, which was to be emphasized by the vertical and symmetric arrangement of subgraphs, is, in actuality, masked by the closure implied by the stretched dollar sign. The graph is mentally encoded as a jagged circle rather than as a pair of diverging lines, which is the encoding that maximizes the chances of noticing the trends of interest. The net result of the misapplied closure here is suboptimal graphic communication.

Often a set of Gestalt principles work in tandem to affect the interpretation of graphs. For example, in a scatterplot, a set of points representing paired observations of correlated variables are seen as an elongated, diagonally-oriented cloud whereas when the variables are uncorrelated a diffuse swarm is seen. The perception of the elongated cloud is presumably caused by the principles of proximity (the points are closer to one another on the average when variables are correlated), good continuation (going from left-to-right, individual points continue the trend of the previous points), and good form (the resulting cloud is relatively compact and has a smooth envelope). The ability to read scatterplots accurately can thus be manipulated by manipulations that affect these Gestalt laws. Wainer & Thissen (1979) show that people are generally accurate in judging correlations from scatterplots when the size of the cloud is not varied (as was done by Cleveland et. al. to illicit an illusion), including the ability to detect and compensate for "outlier" points. This is exactly what the Gestalt laws would predict, because outliers would be perceived as not being part of the cloud (they do not group according to proximity, good continuation, and good form).

INSERT FIGURE 3.38 HERE

A corollary: "Goodness" of Parts

Not only do the Gestalt laws define the perceived relation between one part and another, they also define the perceived relations between parts and wholes. Instinctively, some parts of a pattern are "better" or easier to see than others (e.g., the triangles in a Star of David vs. the overlapping parallelograms), and the Gestalt laws give a more precise definition to this notion of "goodness" of a part (Wertheimer, 1934). Specifically, a part will be "good" to the extent that 1) its own subparts are linked together by the Gestalt laws, and 2) the subparts do not link up with other subparts composing the rest of the figure. The overlapping triangles in a Star of David meet these criteria, but the parallelogram in the middle does not. That is, the subparts of the triangles are grouped according to the Law of Good Form, and these parts do not naturally group with other subparts. The parallelogram, on the other hand, is composed of subparts that are more naturally grouped into other parts (the two triangles) by the Law of Good Continuation. Thus, triangles are good, easily- seen parts, whereas the parallelogram is a bad, hard-to-see part. As we shall see when we turn to processing limitations, the "goodness" of parts will prove to be an important determinant of how easy it is to extract various types of information (e.g., simple values, differences, trends) from various sorts of displays (e.g., tables, bar graphs, line graphs).

2. Integral and Separable Dimensions of Visual Perception

The foregoing principles all determine how separate marks will be grouped together. There is another kind of principle that not only sometimes determines how separate lines will be grouped, but determines which dimensions (such as hue, saturation, intensity, height, and width) will be grouped together. That is, there are cases where we cannot help but attend to one thing given that we are attending to another. These sorts of dimensions are called integral. According to Garner (1970), "Two dimensions are said to be integral if, in order for a level on one dimension to be realized, a dimensional level must be specified for the other." If the specification of one dimension in no way influences the specification of the other dimension, then the two are separable. When it comes to perceiving values on dimensions, however, some dimensions act as if they were integral (i.e., one cannot notice one without the other) even though geometrically it is possible to specify values on one independent of the other. Thus, it becomes a matter of psychology, not geometry, to determine which dimensions bind into single units.

Integrality/separability becomes important to the study of graph perception when we consider that graphs usually convey quantitative information by depicting elements that can be specified in terms of their values on a number of dimensions (e.g., when a quantity is communicated via the length of a bar, the width of a circle sector, the darkness of a patch, etc.). To extract the meaning of a graph, one must mentally "describe" its elements in terms of their meaning-bearing dimensions and then translate that visual description into appropriate conceptual "message" (see Chapter 6). Thus it is crucial that the reader mentally encode the graphic elements in terms of dimensions that can, in fact, be translated into the quantitative variables being communicated, for example, the width of a sector in a pie chart and not its angular position; the heights of bars in a bar graph with a nominal scale and not the degree of curvature of the contour formed by their tops; the relative slopes of lines in a line graph and not their relative lightness. The study of the integrality and separability of perceptual dimensions tells us to what extent these perceptual encodings are possible; that is, whether one can ignore the position of a sector while attending to its width, and so on. If not, the reader will be forced to keep in mind various attributes of the graph that have no communicative function and just consume precious short-term memory capacity.

INSERT FIGURE 3.39 HERE

Two experimental techniques are often used to determine if two dimensions are separable or integral. In one, subjects are asked to sort a deck of cards into two piles. On each card is a two-dimensional stimulus, and each stimulus can have values on two dimensions (e.g., a circle varying in size with a radius varying in angular position; geometric forms that vary in both height and width). Figure 3.39 illustrates a pertinent example, a simple stimulus with two bars, each of which can vary in height. The basic card sorting task requires the subject to sort the cards on the basis of the value of only one of the dimensions, such as the height of the left bar in the stimuli of Figure 3.39. So, all cards with a short left bar would go in one pile, and all cards with a tall left bar in the other. Now, the values of the irrelevant dimension (the height of the right bar) are systematically varied: either they are constant (i.e., they are always tall), or they are independent of the values of the other dimension (i.e., they are either short or tall when the left ones are short) or they are redundant, reinforcing those values (i.e., they are always tall when the left bar is tall and vice-versa). If the dimensions are integral, and here is the key prediction, then whether the two dimensions are constant, correlated, or independent will affect how easily the subject sorts on the basis of one of the dimensions. So, if the pair of bars are in fact processed integrally as a single unit, then the relation between them cannot be ignored: sorting the cards according to the height of the left bar will be easier when the height of the right bar is correlated with it, compared with when the right bar is constant. The subject cannot help but pay attention to the height of the irrelevant bar, and its value being correlated with the variable of interest will improve sorting speed and accuracy. Similarly, with integral dimensions, the independence of the values of the two dimensions will hinder speed and accuracy compared to when the irrelevant one is constant: the subject will be continuously "distracted" by the values on irrelevant dimension, even though in this condition it provides no useful information. On the other hand, if the bars are separable, then the relation between them will be irrelevant, and it will make no difference how the right bar varies when people are asked to sort cards with a short left bar in one pile and a high left bar in the other.

The second technique used to discover which dimensions are integral and which are separable rely on "multidimensional scaling" of stimuli. In these studies subjects are first asked to assess the similarity of each member of a set of stimuli with each other member. These similarity data are then used as inputs to one of a number of standard multidimensional scaling programs (see Kruskal, 1964). In the outputs of these analyses, each stimulus is represented as a point in a space and the distances among the points are proportional (or as close to proportional as possible) to the subjects' similarity judgements, with more similar items (usually) being closer together in the space. Now, there are different ways for the program to measure "distance between points", different distance metrics, when it creates the spatial model conforming to the similarity ratings. "As the crow flies" and "as walking along city blocks" are two examples of measuring the distance between two buildings in a city. And, in fact, these two metrics, called Euclidean and City Block, respectively, have proven important for understanding how different dimensions are processed together in perception. When people assess the similarity of stimuli composed of integral dimensions, a Euclidean metric allows one to construct a better spatial model representing the similarities; in contrast, if stimuli are composed of separable dimensions, a city block metric does a better job. This relationship is a consequence of the mathematics of how the metrics are computed. Intuitively, if the dimensions are separable, then their effects will simply add. And in computing city block distances one merely adds the "legs of the triangle," specifying distances along the x and y coordinates of the space (let us assume it is two-dimensional for this example). If they are integral, however, one "takes the hypotenuse," or "crow's flight", which combines non-additively the individual contributions of the two dimensions. For a technical description of the underlying mathematics, the interested reader is referred to Attneave (1950), Shepard (1964) and Torgerson (1958). It is comforting that this very different experimental techniques leads to the same conclusions as Garner's card sorting task.

A number of investigations have used both the basic card-sorting technique and multidimensional scaling with a wide variety of different dimensions, most of which could be incorporated in a chart or graph. Table 3.14 summarizes these findings, indicating which dimensions seem to be grouped into integral units; the table also provides an illustration of each kind of dimension.

INSERT TABLE 3.14 HERE

If the elements in a graph vary along a single dimension, dimensional integrality will play no significant role in comprehensibility. But we have a strong prediction when several dimensions vary at the same time: if the dimensions are integral, covarying them will make it easier for the reader to encode the information they convey, but if they vary independently, the information will be harder to interpret. For separable dimensions, the degree of correlation will not affect comprehensibility as much. For example, the height and width of rectangles are integral dimensions. If the height of rectangles convey, say, oil reserves of a country, then varying their widths simultaneously will increase the impression of whatever differences may exist, as is illustrated in Figures 3.40a and b. But if the widths vary on their own, as in Figure

3.40c, where width might signal, say, population size, then differences in the first variable will be harder to detect--because the width cannot be ignored when height is attended to, and intrudes into short-term memory. For separable dimensions (e.g., rectangle height and the curvature of the lines filling them), this effect is not as apparent, as is illustrated in Figure 3.40d.

INSERT FIGURE 3.40 HERE

III. Processing Priorities and Limitations

Not all stimuli are given equal treatment; such are life's injustices. Some stimuli are inherently more "salient" than others, and as such grab one's attention at the outset. The factors that determine stimulus "salience," then, will also determine what it is that one is likely to notice and remember about a graphic display. These factors are especially important in light of the fact that one will not remember everything in a display. The virtual inevitability of imperfect recall is in part a consequence of limitations on our initial processing of a stimulus; in particular, by limitations on the "span of apprehension" and the amount that can be held in short-term memory at once. Thus, in this section we will consider two kinds of principles, those pertaining to stimulus salience and those pertaining to limitations on short-term stimulus encoding.

A. The Principle of Stimulus Salience

Under certain conditions, the perceptual properties inherent in a visual display will determine the likelihood and order that a given part is noticed and remembered. By varying the "salience" of the marks, one can facilitate the correct interpretation of the information content if the most salient marks draw the reader's attention to the most important part of the display. The principle of stimulus salience, then, exhorts one to vary the marks such that the right information is likely to be encoded. What constitutes the "right information" is, of course, up to the chart or graph maker to decide. But what properties of marks dictate the priority of encoding? This. unfortunately, cannot be answered definitively in the general case. No single dimension, such as color, necessarily takes precedence over other dimensions, such as shape. For example, consider two shapes that are colored differently. If the shapes are a circle and a very circular elipse, the difference in colors will probably be more noticeable. But if the two shapes are a circle and a triangle and the colors are barely different shades of orange, then the shape will be the more salient dimension. Thus, the chart or graph maker must use his or her own intuitions about what the salient dimensions are in a given display--which is not particularly difficult, once one is alerted to the role that such perceptual "saliency" plays in emphasizing particular aspects of a display over others. In fact, in general one can expect large differences between values (e.g. size, brightness, color) of objects, and extreme values of a single dimension, to be perceptually salient. The human perceptual system is often characterized as a "difference-detector" or "variation-detector" rather than as a detector of steady states or constant stimulation (Lindsay & Norman, 1972; Helson, 1964), so we can expect that discrepancies and differences, especially extreme differences, among stimulus values will capture the reader's attention and find their vary into his or her encoding of the visual aspects of the graph.

INSERT FIGURE 3.41 HERE

For example, consider the graph in Figure 3.41a, which displays a pair of parallel lines. Under normal circumstances the lines will be interpreted to indicate that A is greater than B, with A and B being in a simple relation to each other. Suppose, however, that we want to emphasize that B is less than A (e.g., that the level of US military readiness is less than that of the USSR).

In this case, we want A to be the baseline and B to be defined in relation to it. If we assume that the line noticed first will serve as the "assertion" (and this should really be tested empirically), then simply by varying the weight of the line we can vary which serves as the baseline. Consider Figure 3.41b; now the fact that B is less than A seems to jump out, one notices B first and then its relation to A.

INSERT FIGURE 3.42 HERE

As an example of how the Principle of Stimulus Saliency can lead to impaired interpretation, consider the graphic display of Figure 3.42a. This display illustrates the monthly total flow in two drainage basins. The background is more varied and complex than the information-bearing components of the display, and seems to draw one's attention from those components. Simply by making the lines delineating the background finer, and those comprising the display proper bolder, we greatly improved the legibility of the graph--as is evident in Figure 3.42b.

B. Principle of Finite Capacity

This principle has two parts, one pertaining to the limitations of short-term memory and one pertaining to the limitations of re-organizational processes used during encoding.

1. Short-term Memory Capacity

As a general principle, Less is Best: Human beings can only hold in mind a total of about four units at once, and hence should not be required to do more than this in order to comprehend a graphic display. There is a long history of study of the limitations of short-term memory, which has led to a variety of conclusions. Everyone agrees that memory is severely limited, but the question has been one of how much so. The problem in measuring memory limitations is that the correct unit is not necessary determined by any simple measure of the number of stimulus elements. For example, if one gave someone a list of digits to remember, the person could reorganize the list into pairs of digits (e.g., "twenty-one" rather than "two" and "one"). In this case, there would be half as many "psychological units" in the mind of this subject as in the initial set (as conceived of by an experimenter, who considered each digit as a unit). And there is nothing from stopping the subject from organizing the digits into groups of three or more. Thus, it is critical to determine how elementary units are organized into each "chunk" (the technical name of a set of information that is held in memory as a unit). The classic paper on the subject, by George Miller, lays out one hypothesis in the title: "The magical number seven, plus or minus two: some limits on our capacity for processing information." However, we fear that this number is more magical than accurate. This number is suspect because subjects may have organized stimulus elements into fewer psychological units. In the course of studying the process of organizing elements into units, Ericsson, Chase and Faloon (1980) provide support for a different estimate.

Ericsson et. al. asked one subject to return to the laboratory five times a week over the course of nine months. At each session the subject was given a set of digits to recall. Amazingly, the subject gradually built up to the point where he could remember 79 digits! The digits had to be presented rather slowly, however (one every 5 seconds) to allow the subject time to organize them into units. The nature of these units changed over time. The critical observation Ericsson et. al. made concerned sudden jumps in performance, which occurred when the subject discovered a new, more efficient way of grouping. Critically, at each jump, the number of digits retained was some multiple of 4. At first, the subject remembered pairs of digits, producing a span of roughly 8. At the end, he was able to form groups of 20, and was able to describe the grouping strategies he developed, which perfectly predicted digit span of 4 units were in fact retained. This estimate turns a sow's ear into a silk purse: the grouping strategy which previously had obscurred the number of units being stored now was used to implicate that very factor. And the answer was almost precisely 4.

A number of studies of charts and graphs per se have documented the effect of requiring the reader to keep too much information in memory at once. Washburn (1927) and Vernon (1952) found decreased accuracy in answering questions about a graph as the amount of information that had to be remembered increased. Perhaps more interestingly, Schutz (1961) found that it was better to plot several lines within a single framework than to plot them on multiple frameworks _if_ one was asked to compare values or trends. In contrast, if one was asked merely to retrieve single values, the way the functions were plotted made no differences. This is not surprising, given that only when comparisons are required need one remember where along a function one must make comparison. In this case, having the lines one above the other saves one the effort of remembering the location _and_ the result of making the comparison for each function: now one need only move one's eye up the page, holding a minimal amount of information in mind at once.

The use of keys and legends often will violate the principle of finite capacity. A key is equivalent to a "paired associates" task, where a person is asked to memorize an association between two stimuli. In this case, one must memorize the pictorial information that indexes the different functions (dotted lines, different colored lines, etc.) and the label, and then must match the line segment in the key with the proper specifier element in the graph. In contrast, if one labels the functions directly, there is no need to perform this memory task--which is a boon even if the amount of material in the key does not tax memory. Compare Figures 3.43a and b to see what we mean. Thus, it is not surprising that Culbertson and Powers (1959) found that labels and pictorial symbols that are directly associated with a function are easier to read than keys. Not only does this system save memory effort, but the close proximity of the label and function serves to group them into a single unit (via the Gestalt Law of Proximity discussed earlier), eliminating extra processing required to look up which function is referred to by a label in a key. Similar results have been obtained by Parkin (1981).

INSERT FIGURE 3.43 HERE

One last study of capacity limits in graph comprehension must be noted. Price, Martuza, and Crouse (1974) investigated whether subjects encode point and/or slope information when they are instructed to learn the information presented in a line graph. The authors conclude from their results that subjects encode datapoint information and *not* slope information, and that in order to answer questions about slope, subjects recalled point information and then inferred the slope. This conclusion was supported by the finding that as the number of data points needed to respond correctly increased, performance decreased. This last result is just as expected from the principle of finite capacity under the assumption that subjects do in fact encode point information. Why they would do this instead of encoding slope per se, which involves fewer chunks or units and hence less demands on memory, cannot be stated with certainty. Both the specific instructions the subjects were given initially and the specific questions they were asked could possibly have biased the encoding strategy, but we cannot know for certain because neither are described in detail in the paper. If, in fact, subjects can be "set" to encode graphic displays in different ways, it will be very important to study the textual context in which a graph is placed. This clearly is an area begging for systematic study.

A straightforward consequence of the principle of finite capacity and the principle of stimulus saliency can be stated as follows: A chart or graph should not convey more or less salient information than is necessary for the purposes for which it was constructed. In the ideal case only the information necessary to extract the intended message should be included as salient marks in the display. This pertains to two considerations: how much information is included in a graph itself, and how much non-information bearing visual material (e.g., pictorial backdrops) is drawn together with the graph. As for the former, Tufte (1977) and Wainer (1977, 1978) caution against the temptation to make a graph serve as an archive for large amounts of data in circumstances where the communication of some part of the data is intended: in such cases the unnecessary material can overload the reader's capacity to the point where the intended message is inaccessible. To take a simple example, Figure 3.44a, which is intended to illustrate the Yerkes-Dodson Law (which states that performance will be more accurate with intermediate levels of arousal than in very high or very low arousal). As a general illustration of the Law, the idealization in Figure 3.44a is to be preferred over Figure 44b, which presents too much detail. However, in some cases complexity is unavoidable: if one wants to know the additional details, for example, differences between men and women, then the idealization in Figure 3.44a is inappropriate and the more complex Figure 3.44b is appropriate.

INSERT FIGURE 3.44 HERE

The second practise that can make graphs too complex is the inclusion of superflous salient material in a chart or graph. Tufte (1978) and Wainer & Thissen (1981) criticize the unthinking use of what Tufte calls "chart junk", and urge that the "data/ink ratio" (the ratio of ink used to convey information and ink used for decorative purposes" be as close to 1.0 as possible. As a simple example of the capacity demands of superfluous graphic decoration, compare the two graphs presented in Figure 3.45. The leftmost one contains a set of background elements that vary in size and position, thereby becoming perceptually salient and seeming, on first glance, to convey information. But in fact these figures are used in an attempt to make the display more attractive or interesting, and convey no information in their own right. The middle graph simply removes these distracting elements, making the relevant information more easily seen. On the far right is a graph that retains the decorative elements present in the initial ones, but now reduces the problem of interpretation by making the figures lighter than the information-bearing components (and thus taking advantage of the encoding priorities of the visual system and of grouping by similarity), and keeping the size and orientation of the background elements constant--and hence reducing their salience, their encoding likelihood, and the impression that they convey information.

INSERT FIGURES 3.45 AND 3.46 HERE

As another example of cases in which presenting irrelevant information impairs interpreting a chart or graph, consider Figure 3.46a. Here again the background is patterned merely to make the chart or graph more interesting, but the reader cannot know this at first. By simply removing the background, legibility is improved immensely, as is evident in Figure 3.46b. We should add, however, that if a graph maker insists on decorating a graph, the use of the other Principles discussed in this book (e.g., as illustrated in Figure 3.45) will allow him or her to embellish a display without necessarily saddling the reader with the task of sorting the kernels from the graphic chaff.

A final example of how short-term memory limits and graph comprehension concern the duration of transient memories. Peterson & Peterson (1959) showed that unrehearsed information in short-term memory decays within about 20 seconds (see Klatzky, 1975, for a review of related findings, including the controversy over whether it is time per se or the processing of interfering material in the retention interval that causes information loss). A graph maker should not force a reader to retain information necessary to interpret a graph for a long time. An obvious place to keep this in mind is the relative placement of a graph and the portion of the text referring to it. In fact, Whalley and Fleming (1975) have found that when a display is separated too far from the part of the text that discusses the information displayed in it, often the reader will not even look at the display.

2. Comparing Units or Parts of Units

James Pomerantz, now at SUNY Buffalo, used Garner's card-sorting technique to examine how marks (not values on dimensions) cohere into units. He found that sorting time was affected by how separate marks grouped into units. For example, people could sort the two stimuli "((" and "))" on the basis of the left parenthesis faster than they could sort ")(" and "()" in terms of the left parenthesis. The interesting thing here is that the Gestalt Laws of Common Fate and/or Good Figure operate to bound the two elements in each pair into a group, and it is difficult to consider a part independent of the entire unit. "Breaking up" a perceptual unit and seeing one of its parts in isolation can be done, but it taxes our limited processing capacity. The effort involved in seeing a part in isolation is especially extreme when the part is a "bad part" as defined in the section on Gestalt Laws. When a "bad part" (such as the parallelogram in the center of a Star of David) must be attended to, there must be greater allocation of the mental resources that would otherwise be used in maintaining information in short-term memory. And hence, there will be poorer performance in general.

In the case of charts and graphs, it is critical that a small arbitrary segment of a continous line is a "bad" part by our earlier definition. The Law of Good Continuation tends to cause any single segment of a line to be absorbed into the whole. In contrast, a single bar in an array of bars would be a very "good" part because it is differentiated from the other bars by the Laws of Good Form, Proximity, and Continuity. By the same token, however, it should be easier to attend to a line as a whole than to groups of bars as a whole. The line is seen as a single unit--and stored as a single "chunk"--whereas the bars are seen as many units, which would be more difficult to process and store. Therefore, we are led to a prediction: tasks that require reading information about a specific value--and hence attending to a single point--should be easier for bar graphs than for line graphs, whereas tasks that require attending to the entire information set should be easier for line graphs than bar graphs. For example, reporting single values along a function (e.g., the amount of oil produced over time) should be easier with bar graphs, but reporting trends (the rate of increase) should be easier with line graphs. The literature on graph comprehension summarized below bears out this prediction.

One of the classic studies of graph comprehension and use was reported by Washburn (1927). She presented junior high school children with an essay on the economic history of Florence, and embedded in it a body of data that was displayed in different forms to different groups of students (the forms included a prose paragraph, a unit pictograph, a bar graph and a line graph). Subjects studied the paragraph and then answered questions about the data, the questions pertained to the absolute amounts, differences between amounts, and relative increases and decreases. The efficacy of the different formats proved to depend on the type of data to be extracted. If the viewer had to report on the value of one variable given the value of another (in a set of x,y ordered pairs), a table leads to faster and more accurate performance than a graph; this result was also found by Carter (1947a) and Narwrocki (1972). On the other hand, when subjects must report on differences between two values of one variable corresponding to two values on another variable, or when they must compare sets of differences (i.e., trends), bar graphs and line graphs are (respectively) the more effective media.

The fact that line graphs are ineffective if a reader needs to know absolute quantities was demonstrated by Culbertson and Powers (1959). Subjects were required to note and compare specific quantities on various forms of graphic displays. Both horizontal and vertical bar graphs were found superior to line graphs (and there was no difference between the effectiveness of the two types of bar graphs). When either line or bar graphs were segmented they were less effective, which is interesting because the segmentation was arbitrary (segments did not correspond to any meaningful variation along the x axis). Thus, if subjects attended to individual segments as parts--which seems likely because these segments were perceptually "good" parts--they were not attending to the meaningful information-bearing parts. Lastly, when the specifier elements were presented over an inner framework consisting of grid lines, the graph was more effective-- presumably because the grid lines helped to segment the parts of the line or bars that had to be attended to, and linked these parts via proximity and good continuation to the labels on the x axis.

Schultz (1961) provided evidence consistent with our other prediction, that line graphs should be superior when information about trends had to be extracted. He showed subjects line graphs, vertical bar graphs and horizontal bar graphs. The subjects' task was to compare test graphs with a previously-learned set of patterns and rules for naming trends. Subjects were to study the test graph and choose the matching pattern and rule. Line graphs were found to be superior. This study is flawed, however, in that the task may simply have been easier with line graphs because the test patterns had originally been presented in the form of line graphs. (See MacDonald-Ross, 1977, for a more complete survey of the literature on human graph reading).

Thus, we have reasonable support for our claim: it is easier to extract a single value from a bar graph or table than to read it off a line graph, presumably because is that in the latter case one must break up a single perceptual unit, the line, into "non-good" parts. But if the line does not need to be broken into points, it now is more effective--as one would expect given that more information is represented in fewer chunks (assuming a line is one chunk, as is each bar in a graph or entry in a table). So, reading a trend is easy with a line graph because the information is inherent in the single unit that can be looked up as such. In contrast, extracting general trend information is harder with other media, where the trend must be computed, keeping a number of chunks in mind at once. Note that if precise differences are required, however, how bar graphs are best--here one must extract precise value first, which is difficult with line graphs. In short, then, as long as one does not need to decompose a perceptual unit into smaller parts that are not "good" parts according to the Gestalt laws, then it seems safe to say that the fewer units one uses in displaying information, the better.

CHAPTER 4: SEMANTIC, FORMAL, AND PRAGMATIC PRINCIPLES

I. Semantic principles

   1. Surface compatibility

      a) Typicality

      b) Congruence

      c) Cultural convention

   2. Schema availability

      a) Concept availability

      b) Graph schema availability

II. Formal principles

   1. External Mapping

   2. Internal Mapping

III. Pragmatic principles

   1. Invited inference

   2. Context

CHAPTER 4: SEMANTIC, FORMAL, AND PRAGMATIC PRINCIPLES

In the previous chapters we have reviewed syntactic principles, which are concerned with how our visual systems interpret marks on a page. These principles are content-free in that they operate independently of what the lines mean. Even if none of the syntactic principles is violated, and hence one can detect the marks, read them without distortion, organize them correctly and hold the relevant information in mind at once, a chart or graph may still be defective at a semantic level of analysis. The semantic analysis assigns meanings to the elements and the relations among them. If the wrong semantic interpretation is assigned to a given mark, the chart or graph obviously will not communicate effectively.

Similarly, the mapping from mark to mark in a display may be faulty, or the wrong inferences may tend to be drawn. In such cases, a formal or pragmatic principle has been violated. In this chapter, then, we will consider semantic, formal, and pragmatic principles.

Ultimately, all the changes one makes to improve a chart or graph will be made at the level of syntax. Once a violation of one of the semantic, formal, or pragmatic principles has been detected, it can be rectified by altering the lines themselves. But one should be careful to distinguish violations at the level of syntax proper, such as those discussed previously, with semantic, formal or pragmatic violations. These latter violations only come to light when the chart or graph is considered in its role of a communicator of specific information in a specific context, as discussed below.

I. Semantic Principles

We have formulated two general semantic principles. The first is concerned with the compatibility between the mark used to convey information and the intended meaning. Some symbols are better suited for a given role than others.

These first principles, "The principles of surface compatibility," have three distinct aspects, as discussed below. The second principles are concerned with the kind of knowledge one must have in order to understand a concept. These "principles of schema availability" have three aspects, ranging from the availability of single concepts to individual differences in the availability of knowledge of specific graph types.

1. The Principles of Surface Compatibility

The basic message of these principles is straightforward. The format of a display should be compatible with its spontaneous interpretation. If a mark is spontaneously described in a way incompatible with what it represents, the graph maker is in trouble. This principle has three major aspects, which are implicated in the literature summarized below.

a. Typicality

In a series of very important and ingenious experiments, Eleanor Rosch reported findings that are relevant to how graphic displays should be labeled (see Rosch, 1978, for a more detailed summary). Rosch distinguishes between a "horizontal" and a "vertical" level of classification in a taxonomic hierarchy. For example, take the familiar hierarchy of the animal kingdom, where each beast is a member of a species, a genus, a family, an order and so on. Within a given level, say species, some instances are more _typical_ or "representative" of the category than others. For example, a collie is a more typical dog than is a pekinese. This kind of variation defines the _horizontal_, within-category dimensions. In addition, any given example can of course be assigned a classification at numerous levels of hierarchy. The collie is not only a collie, but a canine, mammal and animal as well. This kind of variation defines the _vertical_ dimension. Just as there is a best example of any given category, there is also a "best category" for any given example. When we see a dog we spontaneously classify it to ourselves as a dog, not a mammal. The level at which we spontaneously classify a typical object is called the *basic level*. The basic level is that at which the examples are as similar as possible while the category itself is as general as possible. So, to take another example, the basic level category for a MacIntosh apple would be apple, not fruit or MacIntosh apple. The reason is that although "fruit" is a more general category, the examples within the category are rather dissimilar (watermelons and tomatoes don't have much in common). "MacIntosh apple," on the other hand, is a category with very similar members, but not much more similar than the more general category "apple." Interestingly, the horizontal and vertical dimensions mutually effect one another: atypical examples, such as a penguin, are *not* named at the basic level. Rather, they are spontaneously named at the most specific level (Jolicoeur, Gluck, & Kosslyn, submitted).

The relevance of this work on categorization is clear when depictive symbols are used in a chart or graph, either as bar elements (e.g., in a pictograph), background figures, or as labels. First, typical members of the category always should be used. "Birds" are best symbolized by robins, not penguins. Second, one should avoid a picture whose "basic level" differs from the level that is the subject of the communicated message. Rakes should not be pictured if the picture is to stand for tools, since the basic level for a specific rake is "rake", not "tool", "object", or "leaf rake", so "rake" is what the reader is likely to think when he or she sees the picture. To symbolize "tool", use some object whose basic level *is* appropriate for tools--such as a tool box, which will probably be encoded as such and not as a "box" or "electrician's tool box". The rules for determining in advance how a picture will be named are complicated and not yet totally worked out. The artist should merely show a depiction to a couple of people (who are representative of the intended audience) and ask for its name; if the name spontaneously given is not correct, the drawing must be altered.

Consider some more concrete examples: Figures 4.1a and b show alternative presentations of the same information. But in Figure 4.1a the reader can be mislead into thinking that the graph is about cars--but in fact the graph is about the rising price of gas, which is clearly evident in Figure 4.1b. In Figure 4.1a, the basic level of the framework is "car," not "oil-burning vehicle", and this conflicts with its role as a constituent in a graph about gas. Figure 4.2 provides another example of how this principle can be used to enhance graphic communication or, if violated, can impede it. The specifier in Figure 4.2a consists of pictures of different kinds of trees, the heights of which represent how much that sort of tree grows when soil is treated with sewage. The immediate meaning of the "bars" in this graph is the very kind of tree being represented, which serves to reinforce the message of the graph. Now consider Figure 4.2b, in which barrels (presumably of sewage) are used instead of trees. Now one is set to wondering about different amounts of sewage, which is not the point (or even indicated) in the graph. Even though the basic level of each tree is "tree", not "tamarack" or "pine", as one would wish in this case, "tree" is closer to "tamarack" than "barrel" is in a person's mental dictionary, and hence the various trees would come to the reader's mind sooner in the first case than the second.

INSERT FIGURES 4.1 AND 4.2 HERE

b. Congruence

This aspect of the principle of surface compatibility has four parts, all of which deal with setting up a "natural" correspondence between stimulus properties and the information they convey.

Cognitive compatibility. Perhaps the most basic form of cognitive compatibility occurs when one makes sure that the physical characteristics of a mark, particularly its size and color, are appropriate for the information one wants to convey. The description of the marks themselves should not contradict their meaning. As an extreme case of mis-alignment, consider the Stroop phenomenon: If I present you with the word "red" printed in blue ink, and ask you to name the color of the ink, you will experience interference. The meaning of the marks conflicts with the color itself. A large mark will be described as large, and hence should not be used to represent something small. Similarly, one should not use small font to spell out the word "elephant" and large font for the word "fly" or complex lettering for the word "simplicity".

INSERT FIGURE 4.3 HERE

Figure 4.3a provides a somewhat subtle, but nevertheless troublesome, violation of the principle of surface compatibility. In this figure we see three groups of three bar elements. Each group refers to a particular output measure in a medical experiment and each bar within a group indicates the output level achieved by one of these types of treatments. According to the vertical axes, the unit being measured is the percent deviation from a base value achieved by an untreated group of animals. However, the physical base of the bars is clearly the horizontal axis of the graph--which represents minus infinity! In order to interpret the information being represented the reader must pay attention to the empty space between the bars when they are less than the baseline, and a relatively small part of the bars when they extend above the baseline. This format clearly conflicts with the concepts of "a little better, a little worse" which are being represented (relative to the untreated controls). Consider how much more obvious are the results when they are graphed as in Figure 4.3b. In this figure, "better" and "worse" correspond in a simple way to a simple relation relative to the baseline.

Many of the violations of this principle involve color. Figure 4.4 presents a common use of color in charts. Different colors are used to stand for different proportions of households with pet fish. Most people have trouble in reading these graphs, however, because different colors do not align themselves into a single dimension to the eye. These kinds of dimensions--in which the values differ qualitatively--can be contrasted with others, such as loudness, where the values differ quantitatively. Red is not "less blue", whereas 100db is less loud than 200db. Qualitative stimulus dimensions, such as color, should not be used to represent quantitative conceptual dimensions. Figure 4.4b presents the same data using degree of shading as the differentiation--note how much easier it is to compare the different regions in terms of relative numbers of pet fish (see also Wainer & Francolini, 1980, who show that maps using transitions from one hue to another to illustrate a continuous variable are difficult to understand).

INSERT FIGURE 4.4 HERE

An exception to this rule sometimes occurs for isolated parts of qualitative dimensions--Guinor and Stevens (1967), for example, found that green, blue and violet ordered quite naturally into a continua, as did red, orange and yellow--although the two sub-continua themselves do not naturally align in terms of a progression from "less" to "more".

Other psychological principles that bear on cognitive compatibility can be gleaned from the literature on synesthesia and cross-modality matching (Marks, 1982). Synesthesia occurs when perception in one sensory modality is accompanied by sensory experience in another. For example, many people report "seeing" colors appropriate to the music they are hearing. The pairing of colors and tones is not arbitrary (e.g., low tones go with blue, high with yellow), and the pairing found in synesthesia is also found in cross-modal matching. Cross modal matching involves a person selecting values along one dimension, such as color, to be paired with values along another, such as pitch. People can pair cross-modal experiences very reliably, including somewhat bizarre combinations such as beer taste with pitch! In addition, people are near unanimous in judging that the vowel "a" (as in "bake") sounds or feels more yellowish than the vowel "o" (as in "not"), which feels more brownish or blackish. One would therefore expect that visual dimensions that connote other sensory dimensions would make better symbols in graphs for those dimensions than other visual continua would. The only empirical test of this prediction, Cuff's (1973) experiments on temperature maps, was taken to disconfirm the prediction; he found that blue was as effective as red in conveying mean temperature of regions on a map, and that a blue-red continuum was less effective than blue alone or red alone (see also Wainer & Francolini, 1980). However, this may have nothing to do with the relative effectiveness of red and blue to symbolize *temperature*--it may simply reflect the ease of perceiving red and blue as lying along a continuum, as noted above (recall that colors vary qualitatively, not quantitatively). Similar effects probably occur with whatever this color scheme is used to symbolize.

In addition, although most synesthesia and cross-modality matching research examines the compatibility of one sensory dimension with another sensory dimension, it seems likely that certain abstract dimensions may "look better" when visually represented one way rather than another. For example, our intuition is that the military strengths of nations are represented well by different sizes on a map or different thicknesses of borders if defensive strength is emphasized, whereas it would be less natural to represent average life expectances of the nations in those ways.

Naming space. There is considerable evidence that there are some general principles of how we conceptualize visual space, and these general principles can be of use to the graph maker. Our preference for conceptualizing visual scenes purportedly arise from fundamental constraints on how we conceptualize actual visual space in the real world (see Clark, 1973; Clark, Carpenter & Just, 1973). Specifically, the important dimensions of physical space are those relative to the observer, namely position relative to ground level and relative to the scope of vision. Even though "up" and "down" are equivalent in terms of the information they convey ("up" can signify "not down", and vice-versa), and the same is true for "front" and "back", psychologically there is a preference for coding the locations of objects in a visual scene as positive if they are up and in front, and as negative otherwise. This makes sense considering that things that are down or behind can become invisible (underground or behind the head) in the real world. There is also a preference to make comparisons in terms of the "unmarked" member of a pair of polar adjectives, which is the member that serves as the name of the dimension. That is, to ask "how high (or tall) is X" does not imply that either is high--whereas asking "how low (or short) is X" implys that both are low. The use of a term which does not label a dimension proper, then, will lead the reader to infer that the variations along the dimension fall on one (usually the "low") end. The brief implication for graph design concerns the labelling of x's and other graphic elements and how they will be interpreted by the reader. If a dimension is labelled as shortness", "smallness", "farness", etc., the reader will draw conclusions that would not be drawn if the labels read "tallness", "nearness", and so on. Furthermore, comparisons of quantities using words like "shorter than", "smaller than", "farther than", and so on, will be harder to understand that the equivalent comparisons using "taller than", "larger than", "nearer than", and so on (Clark & Clark, 1968; Clark & Card, 1969).

Conceptual alignment. Once one has chosen a visual dimension to represent a conceptual variable, how does one decide which way to "align" the two scales? For example, if the oil production of nations is to be represented by the light-dark dimension, should the nations with more oil be colored lighter or darker than the nations with less oil? Cooper and Ross (1975, see also Pinker and Birdsong, 1980) have proposed a rationale for deciding, based on the linguistic phenomenon known as "freezing". Most languages permit conjoined words in idioms to be spoken only in one order, for example, "here and there", but not "there and here"; "kit and caboodle," but not "caboodle and kit". Cooper and Ross outline a set of quasi-universal phonological and semantic principles, experimentally validated by Pinker & Birdsong (1979) that govern these "frozen orderings". Of concern here is their suggestion that the first terms of frozen conjoined phrases are all more "psychologically central" than the second terms, and vice versa. This provides a large set of simple predictions of the "best" way to fix the endpoint of a mental variable with the endpoint of a visual one. Since we say "more or less" and we also say "light and dark", "thick and thin", and "up and down" (but not vice versa in either case), one can link up the first terms of each phrase and the second terms of each phrase. This leads one to assume that more oil should be represented by lighter shades, thicker boundaries, or taller bars; and less oil by darker shades, thinner boundaries, and shorter bars according to Cooper and Ross.

At first glance, this appears contradictory to one's intuitions, which would lead one to make an area darker to represent more oil (or generally more of anything). It also seems contrary to the results of Cuff (1973). Recall that Cuff investigated the differences between qualitative and quantitative methods of shading of colors to convey to readers the desired impression of distribution with the least amount of effort. Children in grades 6 through 12 were shown maps that symbolized the temperature of a given region using 3 color schemes: shades of red, shades of blue, and shades of red and blue. The children were told to consider these maps to be temperature maps, and to mark the areas they considered to be of highest, medium, and lowest temperature. No legends were used, to see whether an effective color scheme conveys the desired information to the reader in a natural way without reference to a legend. Poorest results were obtained with the two color map. It appeared that the qualitative associations of strong red with warm and strong blue with cool (see the forthcoming section on synesthesia and cross model matching) were not enough to override a tendency to associate light shades with low temperatures and vice versa. Deeper shades of blue (as well as deeper shades of red) successfully symbolized warmer temperatures, despite the inappropriateness of the blue as a symbol of warmth, and despite the freezes "light and dark" and "hot and cold", which would make "light" the natural symbol for "hot", one would think.

How can we interpret these anomalous findings? Recall that graphs and maps are almost always shown on white backgrounds, serving as the context against which the marks are defined. Given this, it is natural to interpret the darkest shades as those with the more ink on the page, or the most filled-in, or the most marked. Higher temperatures ("high and low" or "hot and cold") are represented by more ink ("more or less", "filled and unfilled", "shaded and blank"). In support of this conjecture, we have often noticed that lecturers will refer to shaded regions on a blackboard as "dark" or "black" and empty regions as "light" or "white"--even though the opposite is literally the case! Evidently, the "filled-unfilled" dimension is more salient than "light-dark", and this is reflected in people's preferences for dimensional pairing according to the freezing principle. Thus, filled regions should be used to represent the first term in a frozen ordering, and unfilled regions the second term.

c. Consistency with Cultural Convention

A graphic display should not violate common cultural conventions. Much of the way symbols are used is determined by simple convention. But once such conventions are established, it becomes defeating to try to ignore them or, worse yet, fly in the face of them. An effective graphic display, then, will not violate cultural conventions. One such convention is that we normally associate the change with a movement from left to right presumably because that is the way we read. Thus, we are used to interpreting a line that goes up as it progresses from left to right as indicating that some quantity is increasing (in fact, this is even reflected in the way we described the line--note that we did not say that the line went down as it progressed from right to left)[1]. If the same display were rotated 90 degrees, the line would seem to go down, directly violating the interpretation that would follow from the usual convention.

---

[1] Note that this particular case may not be an isolated convention in our culture, and may not be a convention at all. The freezing principles predict just these pairings, since we say "higher and lower", "more and less", "up and down", and "right and left".

Other types of conventions for representing information in charts and graphs are rampant. For example, "red" indicates "stop," not "go" (and vice versa for green), the direction of movement around a circle is clockwise, and so on. Not only are there general conventions in the culture, but each discipline and subculture has its own special conventions. The greek symbol Sigma represents summation to statisticians and engineers, and probably should not be used to represent something else to these readers.

2. Schema Availability

Understanding a chart or graph involves translating a visual pattern into a set of conceptual or quantitative relationships. To do that, the reader must have some general grasp of the conceptual/quantitative relations that the graph is trying to convey, and he or she must know the translation scheme by which the visual marks stand for quantitative information (which will differ from graph type to graph type). We have called this translation scheme a graph schema (see Chapter 6), and whether or not a reader has a graph schema and the concepts it presupposes will affect his or her ease in reading the graph. A good chart or graph, then, should not incorporate concepts the intended readership will have difficulty understanding and should not use formats that are unfamiliar to the intended readership. In this section we consider the data that bear on these principles, and attempt to discover whether there are any data about which kinds of concepts should not be used for given types of people.

a. Concept Availability

No one can understand a graph if he or she does not grasp the concepts themselve that the graph is trying to communicate. One example of a violation of this principle occurs when the graph maker overestimates the sophistication of the readership. For example, consider a graph illustrating the percent change in prices for wheat for each month of the year. The height of each bar represents this change, and hence the entire graph constitutes a plot of the rate of change over the period. If this graph had appeared in an elementary school text, most of the readers probably would not understand it. If the readership does not understand a concept, the information can still be presented. But now the graph maker must use the more elementary concepts upon which the more complex one was built. In this case, "change over time" for different months seems clear enough. But if one were to plot 12 functions on a graph, with the Y axis representing price and the X axis time, the principle of processing limitations would be violated. So, if the point is simply to show that the rate is changing over the year, four plots--one per quarter--would do. The reader need only note the fanning pattern of the functions to get the message. Of course, if the readership is more sophisticated it may be better to use the more sophisticated concept, and save the additional ink.

There is either an embarrassment of riches or a depressing dearth of systematic research on this aspect of the principle of schema availability, depending on how you look at it. On one hand, virtually all the work on children's concepts can be taken to bear on the principle, but on the other hand there has been virtually no work systematically examining what percentage of different segments of the population are comfortable with concepts that conceivably could be used in a graphic display. Some of the work on children's concepts does in fact bear directly on the kinds of concepts necessary to understand charts and graphs. In particular, Piaget (see Flavell, 1963) has provided us with much information about children's mathematical, logical and conceptual competences. Among the relevant research is Piaget's investigation of how and when a child comes to realize that various objects differing along some quantitative dimension can be interrelated by the concept of a scale or a series (see Piaget, 1967). For example, very young children cannot arrange a set of sticks in order of increasing length. If this really reflects a lack of competence for reasoning about scales, then it seems likely that until that awareness develops (around 5-7 years for most domains, according to Piaget) the child will have trouble relating the curve on a graph or an array of bars to the concept it depicts. The conceptual abilities of the child are beyond the scope of this book; here we simply wish to point out the obvious relevance of this research to the problem of what children can be expected to extract from graphs at all, and refer the interested reader to Piaget's books on space, number and merging (Piaget, 1954, 1956; Piaget and Inhelder, 1971; see also Gelman and Gallistel, 1979, for further developments and some cogent critiques of some of Piaget's work in these areas).

Note that there will also be cases in which a reader possesses the concepts involved in a graph's message, but does not realize that the graph is communicating those concepts because the words used for labels are not in that reader's vocabulary. Symbols should be used that will be easily understood by the presumed readership. Most academics, and some people in the business community, seem afflicted with a desire to use unusual words or symbols in place of more familiar ones, often resulting in a violation of the principle of schema availability. If a chart or graph is to be presented in a publication that is directed to a specific, well-defined readership, some jargon may be appropriate to prevent the use of lengthy locutions; however, as a general rule--and especially if the readership is not precisely defined--jargon in labels or special symbols should be avoided.

b. Graph Schema Availability

A person must understand the conventions and notations used in a graphic display in order to comprehend it. If a person has never seen a given graph type, it presents a problem to be solved, at best (if the person is highly motivated), or a road block, at worst. In fact, it has been shown that a very common response to the task of analyzing an unfamiliar type of display is to ignore the display altogether (Wright & Threlfall, 1980). Thus it seems advisable to stick to conventional formats for graphs unless the pattern of information to be communicated is so subtle, unusual, or complex that no such conventional graph can convey the pattern simply (e.g. the novel graphs for the presentation of statistical properties, discussed at length in Tukey (1977) and Wainer & Thissen (1981). Furthermore, it is important to know what sorts of graph types are most effective for what sorts of people. But as we have seen in the previous sections, even this goal must be qualified: the effectiveness of a chart or graph also depends on the uses to which it will be put. Thus, we must take care to evaluate graph effectiveness for different populations performing specific tasks. Researchers in the field have recognized this, and many of the studies in the literature address this topic. Unfortunately, the research generally is so flawed that not much can be inferred from the results. Improvements in methodology and theory have rendered most of the previous work uninterpretable. In this section, we will briefly review the most widely-cited studies in this genre, and will briefly critique them, referring the reader to MacDonald-Ross (1977) for discussion of additional studies. If nothing else, this exercise should be useful to readers who plan to do original research themselves or who plan to read the primary literature in detail.

A. Developmental Research and its Pitfalls

A number of studies have examined how well children of different ages and grades comprehend charts and graphs. Mathews (1924) gave children a representative sampling of graphic materials that pertained to one particular course of study, social science. Various forms of bar, line and circle graphs, time lines, and pictograms were included in the study. The measure of difficulty was the percent of objective questions correctly answered by the members of each group. The results are difficult to interpret, however (as the author himself acknowledged), because there is no way of telling which specific components of the materials were responsible for ease or difficulty of comprehension. The type of graphic components that the author varied included the orientation of bars (horizontal or perpendicular), number of shades on bars (i.e. if each bar represented more than one variable, if they were segmented), number of lines (on line graphs) or divisions (on circle graphs) and the presence or absence of scales. Thus, although Mathews found that the circle graphs were easiest, with bar and line graphs being progressively more difficult, these results are suspicious. Critically, different questions were used with different displays. Thus, the observed differences may have nothing to do with differences in the graph types per se, but only with differences in the difficulty of the various questions. There was no attempt to investigate systematically which types of graphic displays are better suited for deriving which type of information.

Another study that focused on the development of the comprehension of graphs is reported in Strickland's thesis (1938), in which children in grades 1-4 were taught aspects of the history of technology using various sorts of graphs. The question of interest was: which forms of graphs are suitable for each of these four grade levels? A graph was considered "suitable" if children at a given level answered questions at a level of accuracy exceeding chance performance by 30%. She concluded that first graders were "ready" to learn from unit pictographs and from developmental picture charts (a series of pictures depicting some characteristic of successful epochs, such as means of transportation). She also found that second graders are "ready" to learn from circle graphs as well, and so on up the academic ladder. However, given the arbitrariness of her criterion, it is hard to take such conclusions seriously, especially since all grade levels responded above chance accuracy for all types of graphs.

Of more potential interest is Strickland's ranking of different graphs in terms of accuracy of understanding by different grade levels. She found that the ranking of graph types did differ from one grade level to another, which at first glance seems to suggest that there are interesting differences in the development of different cognitive components (as opposed to a simple monotonic increase in attention or other skills, which would have left the ranking intact over grade levels). Unfortunately, different grades were presented with different examples of each type of graph, roughly adjusted to the children's level of skill, thus making direct comparisons of graph types across grades impossible. In fact, the number of correct answers overall did not increase with increasing age--confirming that the corresponding graphs for different grades differed in intrinsic difficulty. Nonetheless, there were some consistent differences among graph types for all ages: line graphs were consistently harder, whereas "developmental picture charts" (series of pictures exemplifying a historical trend) were consistently easier. However, the different formats shown to children depicted different sets of data, preventing us from knowing whether it was the line graphs per se or the particular information set that Strickland chose to depict by line graphs that led to poor performance. Finally, Strickland ranked the kinds of information depicted in graphs in terms of how early children could consistently answer questions about them. She found that as children aged, they became better at answering questions that compared several units of information as opposed to single units, at reporting absolute quantities and precise rates as opposed to relative amounts (i.e., "greater than" and "less than"), at deducing the purpose of the graph, and at reasoning about the information in the graph. Again, the value of these observations is dubious--they could indicate developmental change in attention, interest, memory, reasoning power, perceptual acuity, or a combination of these factors.

Other developmental studies of graph comprehension yielded scattered findings, Vernon (1950) reported only that performance was poor overall; Vezin (1974) found improvements in performance with age, and overall advantages for concrete over abstract material; Malter (1948) found that younger children are poor in recognizing conventional symbols (such as arrows symbolizing movement) in diagrams; Washburn (1927) reported no effects of age among junior high school students; and Zwaga and Boersma (1973) reported a slight advantage of young adults over older adults in recognizing stylized symbols for railroad facilities. Both Vernon (1950) and Strickland (1938) failed to find correlations between children's performance for a type of graph and their accuracy in reporting information from it. Unfortunately, as this heterogeneous collection of studies attests, nothing even resembling a systematic approach examining developmental changes for each cognitive component involved in the comprehension process has been attempted.

B. Individual Differences Research and its Pitfalls

Vernon (1946, 1952) has investigated, in a qualitative way, the retention of material from graphs, testing the hypothesis that graphs will succeed where other methods fail in educating the "man in the street". She presented subjects of varying levels of education with sets of demographic data graphed in various ways and asked them to answer questions or write paragraphs about the graphs. She was impressed by the generally sketchy and inaccurate recall of data, and by the subjects' failure to draw logical, coherent conclusions from the data depicted. She also found that recall increased with the educational level of subjects.

Aside from dispelling the naive notion that graphs are a panacea for ignorance, Vernon's studies are of little interest, since again, they measure the effects of a large number of cognitive processes acting in concert. For one thing, it appears that many of the subjects simply did not understand the questions as the experimenter intended them to be understood. For example, they frequently described the superficial visual appearance of the graph instead of the data contained in it, or they answered questions on public policy according to their own opinions instead of according to the narrow implications strictly suggested by the data in the graph. In addition, subjects had to answer questions from memory instead of responding in the presence of the graphs. Thus, they could have perceived the data from the graph perfectly well, but could have forgotten it rather quickly--a plausible interpretation, given the probable lack of interest in the data on the part of the subjects. In fact, the memory requirement can distort not only the absolute amount of information gathered from a graph, but also the differential effectiveness of different types of graphs, or of graphs as compared to other media. These possible distortions could be due to the fact that different sorts of displays may be encoded in different formats in memory (such as images or words, which may decay at different rates). Also, because of the responses required of the subjects, graph comprehension was confounded with general verbal ability. Finally, the effects Vernon observed of education and intelligence on retention are also of negligible value; the differences that were observed consisted simply of lower retention by the less educated/intelligent subjects. This could betray differences in interest, attention, memory, comprehension of instructions, comprehension of graphs, knowledge, ability to reason about the information conveyed by the graphs (e.g., how to derive the rate of increase in a population knowing the birth and death rates), or some combination of these factors. Further, socio- economic status is confounded with intelligence and education level in this study. Thus, the Vernon experiments are a good demonstration of the pitfalls that can be encountered in testing comprehension of graphs when the different cognitive components are not considered separately. Later in this book we will provide an analysis of these components and make suggestions about how this analysis could guide further research.

There is also a dearth of research on other individual difference variables. Strickland (1938) found no sex differences among subjects, though Vernon (1950) found boys were more accurate than girls, in accord with the large literature suggesting that boys are better than girls at spatial and quantitative reasoning (see Maccoby, 1966). Given the problems with this research, however, even this result must be taken with a grain of salt.

Like the developmental studies, the individual difference studies would be nearly useless even if they had yielded reliable results. "Main effects" of an individual difference variable (i.e., across the board advantages or disadvantages) are basically uninterpretable, since they could reflect differences in a host of variables, such as knowledge, memory, attention, or interest. Once more it must be stressed that the target of such research must be a characterization of differences in operating characteristics of different cognitive components in different populations; for example, boys might be less prone than girls to organizing figures according to the Gestalt Law of Common Fate.

II. Formal Principles

There are many places for potential slips 'twixt cup and lip' in reading a graphic display, and a major one lies in the link between the actual marks and the literal meaning drawn from them. The reader will be attempting to translate each visual element on the page into some conceptual entity or relationship, and this will be difficult or impossible if there is not proper mappings between marks and concepts in the graph itself. We have formulated two formal principles that seem to capture the critical ingredients of a correct mapping and which, if violated, result in an ambiguous or misinterpreted display.

A. The external mapping principle

Every mark should map into one and only one semantic category, and every piece of information necessary to read the intended information should be indicated unambiguously. The first part of this principle corresponds to Goodman's criterion of disjointedness, and the second to his criterion of differentiation (see Chapter 2). This principle will be violated if a mark is ambiguous (such as a specifier bar that could be interpreted as containing two abutting segments or one longer part with a smaller one laying over part of it) or a necessary set of marks is missing (such as numbers demarkating a scale). The important distinction between this source of ambiguity and ambiguity that can arise due to operation of the syntactic principles (especially those pertaining to discriminability and grouping) is that there is no geometric or pictorial transformation of the existing graph that could correct this sort of formal ambiguity. For example, if ambiguity arises because a label is equidistant from two axis, and hence is grouped equally well with each one, this could be corrected simply by moving the mark. But if the mark is inherently ambiguous or missing, no amount of rearranging the existing display will correct matters.

The importance of violations of this principle is context-bound to an unusual degree, as virtually any continuous function or axis in theory violates it (see Chapter 2). In these cases, it is impossible to identify any given location with absolute precision--no matter how one alters the graph (blows it up, etc.). But in virtually all cases, absolute precision is not necessary for the reader to get the intended message. In fact, excess precision will get in the way. For example, marking off 1/100th of a gallon of oil production on a scale is superfluous if the reader is supposed to see how production changes per year--and will tax reader's limited capacity to process information.

Similarly, whenever a picture is used as a label, this principle is technically violated. Any picture can be assigned an infinite number of interpretations, in theory. A picture of a sitting man, for example, could be a picture of a man's head, bent knees, John, a sitting Caucasian, etc. But if the graph maker obeys the elementary principles of symbolization we outlined when discussing the principle of surface compatibility, this should not in fact be a problem: the correct picture *will* be given only a single label by virtually all readers.

Consider the graph shown in Figure 4.5, which shows weight gain for laboratory animals over time for two different kinds of food. The variables represented in this graph are weight, time and experimental conditions. Note that the value of the treatment variable is represented by the level of the dashed line. After week number six, however, the value of the treatment variable changes from "Test" to "Control". This difference in the value of the variable is not represented by a difference in the specifier mark (the line representing the function). Thus, the vertical mapping principle is violated; a meaningful difference in what is represented is not indicated by a difference in the mark.

INSERT FIGURES 4.5 AND 4.6 HERE

This principle is especially important when a reader is supposed to be able to assign values to discrete categories of some kind. If the categories themselves are not represented by discrete marks, this will be difficult, if not impossible. Figure 4.6 illustrates two ways of presenting the same information, the graph on the left violating the present principle. If the reader is supposed to be able to discern which color will be associated with which temperature, the chart on the left is clearly inadequate.

INSERT FIGURE 4.7 HERE

Consider Figure 4.7. This kind of display is common when one wants to display additive components of a set of numbers. Figures 4.7a and b show two ways in which the physical mark "-" can be interpreted. One interpretation is that the mark is a composite of two contiguous marks, labeled in the graph "x" and "y". Another interpretation is that the mark is a composite of two overlapping marks, with "x" included in "y". Note, then, that depending on how the physical mark is interpreted different sorts of information will be inferred. By using the format of 4.7a, it is clear that stopping distances is being assessed only after braking has begun, whereas by using the format of 4.7b, it is clear that the total stopping distance includes the time to begin braking.

Thus, a formal ambiguity requires not just re-arranging or re-scaling parts of a display to correct, but more fundamental changes in how information is presented.

In principle, the most severe violations of the external mapping principle occur when a basic graphic constituent is missing. Figure 4.8 presents some common examples one often sees on blackboards. Without the framework, a person not privy to the coversational context of the graph cannot know the baseline or variation along the relevant scales. But recall that one of the basic ideas of our approach to graphic design is the notion of purpose-specificity: depending on the purpose of a chart or graph, certain information will be required and other information will be superfluous. So even here, if only a trend were required, and a verbal context provided the relevant background, even the quasi-graphs of Figure 4.8 could be adequate.

INSERT FIGURE 4.8 HERE

Having said this much, it is necessary to point out how the external mapping principle is related to the principle of surface compatibility. Some aspects of the principle of surface compatibility, the reader will recall, hinge on marks being interpreted both as a depiction and as a symbol. For example, marks could serve to delineate a framework of a graph on rising gold prices while at the same time depicting a bar of gold. Or, a mark can serve as a specifier (a bar in a bar graph) while also depicting a tree. The vertical mapping principle applies separately to the role of a mark as a symbol and the role of a mark as a depiction. In both cases the interpretation should be unambiguous, and the clear interpretation of the meaning of the symbol does not guarantee the clear interpretation of the depiction and vice versa (e.g., the marks may serve well as a framework but be confusing as a depiction or vice versa).

B. The internal mapping principle

The correspondence between portions of a display should be unambiguous.

The foregoing principle was concerned with the direct interpretation of the meanings of marks in isolation. This principle is concerned with the interpretation of the relations among marks--between specifier elements and labels, between axes or framework constituents and specifier elements, between subgraphs and the main graph, and so on. It is possible to have perfectly interpretable marks for which the interrelations are not clear. Given that all charts and graphs communicate information by displaying some kind of mapping between entities--either between different quantities or different qualities--the necessity for easily-read associations is obvious. And yet, it is very common to discover cases in which the relations among different parts of the display are not clear. For example, consider the graph of figure 4.9a, showing levels of Dow Jones Industrials from 1927 through 1937. The insert represents a magnification of a portion of the display, which is indicated by the bracket. Note that the portion indicated by the bracket does not exceed the 350 level, but the corresponding insert represents a portion extending beyond the 350 level. The puzzlement caused here is clearly eliminated in figure 4.9b.

INSERT FIGURES 4.9 AND 4.10 HERE

Consider now figure 4.10a, which presents a graph used in a textbook on physiology. Here we see six groups of bar elements representing six different physiological and pathological states. Each group is composed of three elements and each of these elements represents a particular property or component of blood. Note, however, that two vertical axes are present at the left of the graph. Each axis is associated with a numerical scale and a label indicating the relevant units. But the association between each axis and the relevant bars is missing: we don't know how to read the meaning of the individual bars. Consider how much easier it is to interpret the chart when the principle is followed and it is correctly labeled, as is illustrated in Figure 4.10b.

The foregoing examples were intended to provide clear illustrations of violations of the internal mapping principle. However, for other types of graphs these violations may be less obvious at first glance. Consider the multiple framework display shown in Figure 4.11, which illustrates receipts for particular services from the years 1976 to 1979. Were the receipts for Business Services in 1979 greater than the receipts for the Hotel/Motel group? If you answer by looking at the specifier, the answer would be "no". However, the two vertical axes do not use the same scale values. Thus, in fact the answer should have been "yes". In this case there is a faulty correspondence between two elements of the frameworks of sub-graphs. Note, however, that this problem is very much bound to the potential use of the graph. If the display were intended only to allow one to compare relative trends over time, then there is no *impediment* (although, technically speaking, a *violation* exists). Figure 4.12a and 4.12b present alternative ways of illustrating this same data which do not fall prey to the violation of this principle. Note, however, that in normalizing the scales other things are lost (such as an ability to read easily the variations among the less profitable businesses).

INSERT FIGURES 4.11 AND 4.12 HERE

So, the message again is clear. Once one is aware of the sources of potential problems with respect to a given purpose, it is usually easy to see how to circumvent them--though the graph still cannot be all things to all readers.

III. Pragmatic Principles

The principles reviewed thus far have been concerned with how charts and graphs convey information as complex symbols. As such, we have considered how the marks on a page are analyzed and grouped by the perceptual system and how the literal meaning of these marks is assigned. But comprehending a chart or graph involves more than merely assigning a literal interpretation to symbols, just as understanding language involves more than interpreting each word and the relations among them literally. As is the case with linguistic utterances, graphic displays occur in a communicative context. The reader is expected to draw inferences and to be sensitive to connotations that are not explicitly present. These indirect statements can either be accurate or misleading, and the graph maker may sometimes intentionally make a "political statement" with a misleading connotation. We will not comment further on the principle that deception is unethical, but will only point out the dynamics of how charts and graphs come to convey information indirectly. Given an understanding of these dynamics the chart or graph maker can be aware of the potentials for inadvertent deception--and the chart or graph reader can be alerted to detecting cases in which he or she may be systematically misled.

We shall consider two general classes of pragmatic principles. The first pertain to the inferences we are invited to make when viewing a display, and the second pertain to the effects of context on how we see and comprehend a display.

1. Invited Inference

If one is asked, "Can you open the door?," one does not say "yes" and leave it at that (pesky thirteen-year olds excepted); rather, one opens the door. But strictly speaking, literal interpretation of the utterance is that a question is being asked. What is happening here is an example of pragmatic factors at work in the comprehension of language. One draws an inference above and beyond the literal meaning. We can outline a number of ways in which charts and graphs can be constructed to induce readers to draw particular inferences. Many of these devices are discussed at length in Huff's excellent book, How to Lie with Statistics. The basic idea of all of these manipulations is the same: use physical properties of the display in such a way that the description of the display itself will exaggerate or downplay specific information.

INSERT FIGURE 4.13 HERE

Labels

The words and phrases used as labels can dramatize a point. In the first panel of Figure 4.13, the title makes an unemphatic statement about the content of the graph. In the second, the word "increases" is replaced by the word "soars", bringing to mind a set of connotations not implied in the first place. Now it is taking off. In the third version the word "inflation" is replaced by the words "runaway inflation". Now it is not only taking off, but dragging everything with it as it careens away! Thus, although the same information is presented in the display, the way it is labeled affects the way it is interpreted. Though these differences may seem trivial, their effects are not. Elizabeth Loftus (1979) has found that when people witness an event, subtle changes in the wording of questions asked afterward have considerable effects in people's recollection of the events. Thus, when asked a question like "How fast was the Ford going when it smashed into the VW?", people give higher estimates of the Ford's speed than people asked the more neutral "How fast was the Ford going when it collided with the VW?". Furthermore, the first group, but not the second, mistakenly "remembered" having seen broken glass in the original event. Little words matter.

Another way labels can be used to give the "wrong" impression is through their absence. As Huff (1954) points out, if one has a small effect and has expanded the vertical scale to amplify it one can conceal this fact through the simple device of failing to label the units and the scale. (Of course, this violates the external mapping principle, but even if the graph were unambiguous, the reader might draw the wrong inferences if the axis labelling was not salient, or if the reader did not have mastery of the conceptual distinction between *absolute difference* and *proportional difference*.

Framework Variations

Scale Units. The use of different scale units is another way in which a reader can be led to draw inferences, and was one of the chief ways to lie with statistics that Huff documented. Because our cognitive systems encode not only the magnitudes of physical continua, but also sort them into a small number of discrete categories in memory (Miller, 1956; Kosslyn, Murphy, Bemesderfer & Feinstein, 1978), once a visual mark is expanded to a certain size, it will be "bumped" into a new perceptual category, and may be represented internally as "tall" rather than "medium" or "4 inches high". And, if Pinker's conjectures in Chapter 6 are correct, people familiar with a given type of graph will translate perceptual categories such as "very dark" directly into conceptual categories like "very large". Thus, a gradual change in a physical continuum such as would be accomplished by stretching an axis, may be encoded in short-term memory as if it were a quantum change, which would then be translated into a quantum change in the quantitative message that the reader carries away. Consider the difference in the apparent increases in figures 4.14a and 4.14b, which vary only in the amount of compression along the vertical axis. In general, when one selects a large scale unit, one is implying that the amount of increase is small; conversely, one amplifies an effect by selecting a larger vertical axis, spreading out the scale units. The use of logarithms instead of linear units can have similar effects.

INSERT FIGURES 4.14 AND 4.15 HERE

Truncation. The way a difference appears can also be manipulated in a graph by truncating the vertical axis and expanding the portion of the scale that remains. The two graphs in Figure 4.15 represent the same information, about numbers of US and USSR missiles. But the one on the left was drawn by a SALT II proponent and the one on the right was drawn by a SALT II opponent. By deciding to begin the scale at 100,000 (a number we just made up, by the way), we could spread out the remainder of the scale--amplifying the apparent difference.

Aspect. One of the most powerful ways of slanting a given graph (if you will forgive the pun), is by altering the aspect of the axes, or ratio of their scales. Figures 4.16a and b show alternative presentations of the same set of data. In Figure 4.16a the ratio of the vertical axis length to the horizontal axis length is 2:1, whereas in Figure 4.16b the ratio is 1:2. Note that the increasing trend in the data is much more striking in the first graph.

INSERT FIGURES 4.16 AND 4.17 HERE

3-D. If a framework is made to project at an angle in space (e.g., it is "painted" on a wall that one examines from an extreme angle), the foreshortening that results can emphasize or de-emphasize a trend. This is because we do not perceive line drawings of extreme perspective projections accurately (Kubovy, in press; Hagen, 1981). For example, in Figure 4.17, widely disseminated by the Reagan administration in 1982, consumer prices are seen to take a noticeable drop. However, the decline appears far less impressive when we consider that the drop is only 0.3% of he CPI, an amount whose tininess is obscured by the fact that it is expanded in perspective in the extreme perspective view of the graph depicted.

INSERT FIGURE 4.17 HERE

Specifier Depictions.

Numerous inferences can be invited by different depictions serving as the specifier in a chart or graph. In a graph that presents the number of annual traffic fatalities over a decade, ordinary bars would suffice to present the data, as is shown in Figure 4.18a. But the implications of those data are really brought home when the bars are replaced with stacks of human skulls, as is evident in Figure 4.18b.

INSERT FIGURE 4.18 HERE

Correlated Variations of Integral Dimensions.

It is possible to create an impression that a trend is increasing more than it in fact is, by taking advantage of the fact that vertical extent, horizontal extent, and extent in depth are integral dimensions. Thus, the value on one dimension cannot help but affect the value on the other. In using a bar graph, then, expanding the width of the bars as their height increases will leave the impression that the increase is more pronounced than it actually is. Similarly, if pictograms are used (in which a picture serves the role of a bar), the size of the entire picture can be varied as well as just the extent along the relevant dimension. In addition, using heavier lines as bars increase in size will underline the increase, as will shading them darker as they become longer. These variations can be combined in any order to calibrate how much distortion there will be in the impression conveyed.

Selective Reporting

Many charts and graphs are idealizations, omitting details that are considered unnecessary for the intended purposes. This principle can be carried to extremes, however, as is illustrated in Figure 4.19. In the left panel is graphed the complete set of data, revealing an inverted U shaped function over time; in the right panel is graphed only part of that function, revealing an increasing trend. If the rightmost graph is correctly labeled as presenting data up to only a specified time, it is literally correct. But as an "idealized" representation of the trend, it is misleading, since the reader will most likely interpret the abscissa of the graph as denoting a representative interval taken from the scale of interest, and hence will falsely conclude that one variable increases with the other in the general case.

INSERT FIGURE 4.19 HERE

2. Contextual Compatibility

Most graphic displays occur in some context, either in text or as part of a discussion. Depending on how the material being graphed is conceptualized prior to seeing the graph, a given display may be more or less appropriate. The message here is simple: a spontaneous description of a chart or graph should not conflict with the description generated on the basis of contextual factors.

a. Compatible Inferences

The connotations of the written or spoken background should be compatible with those of the display. For example, if the text states that "The price of gold soared to $1040 an ounce", the axes of the graph should be constructed such that the function seems to soar. Compare the two graphs in Figure 4.20; which is most compatible with the foregoing statement?

INSERT FIGURE 4.20 HERE

b. Compatible Terminology

The labels in a graphic display should not use different terminology than is used in the text.

c. Compatible Discourse

A graph should not present more or less information than is required for its specific purpose. More information will distract or confuse, and less information will defeat the purpose of the display.

Thus, we now have considered all of the principles gleaned from the psychological literature and generated via our analytic scheme. In the next chapter we will use the scheme to discover the most common kinds of problems with graphic displays.

CHAPTER 5: USING THE ANALYTIC SCHEME: A SHORT FORM AND SOME RESULTS

Our diagnostic scheme as presented in Chapter 2 performed as promised: it reveals problems with charts and graphs in a systematic and well-motivated way. Furthermore, because of its exhaustiveness and attention to detail it helped to induce many of the operating principles discussed thus far. However, although its thoroughness was necessary in the beginning, this characteristic becomes a serious impediment to using the analytic scheme on a routine basis. Clearly a shorter and more directed form of diagnostic instrument is required. This chapter presents a new version which takes the form of a questionnaire. In the following pages we discuss the development of this new version; we highlight its advantages in terms of usefulness to the general graphic practitioner, demonstrate its application to the two graphs which were used to introduce Chapter 1, and finally discuss the results of applying the questionnaire to a substantial and representative sample of charts and graphs.

Development of the Questionnaire

The original scheme is exasperatingly long and must be applied by someone who is thoroughly familiar with the theory developed in the foregoing chapters. In addition, it can only be applied to a graph which is already in existence and only through repeated use on many graphs can one become familiar with the more likely kinds of violations of operating principles. Thus, the original scheme is neither a practical way of analyzing charts and graphs nor is it a very useful tool for teaching one to become a better graph designer.

The motivation for developing a questionnaire format was threefold. First, this format reduces the amount of work, in general, and is easier to use--especially for unpracticed people. One need only attend to the particular areas addressed rather than to construct a complete description of the graph, as was measured by the old scheme. Second, the questionnaire offers a convenient means of summarizing our experience in critiqueing graphs for the reader. The questions it contains were formulated after carefully reviewing many charts and graphs and identifying the most likely ways that particular graphic constituents (i.e., frame, specifier, etc.) will violate each operating principle. The questionnaire is therefore a convenient way of passing along the benefits of our experience to the reader and, as we shall soon demonstrate, also provides an effective tool for troubleshooting existing charts and graphs. Third, since the questions comprising the questionnaire summarize the more common ways that a graph maker can go astray, this instrument itself can serve as a learning aid.

Because the sample of charts and graphs that were used as test cases played such a central role in the development of the questionnaire, the sampling plan by which graphs were selected merits some discussion. The sample had to serve two purposes. First it served as the basis for developing a questionnaire that was applicable to a wide variety of charts and graphs. Second, application of the final version of the questionnaire to a subset of this sample was to yield information on the incidence of operating principle violations of various types in different broad categories of displays. In order to accomplish these goals the sampling scheme first had to be completely independent of our operating principles to insure that further development of these principles was not biased by the sample per se. Second, the sample had adequately to reflect the diversity of charts and graphs with which a reader may come in contact.

INSERT FIGURE 5.1 HERE

The sampling plan is shown in Figure 5.1. Note that this plan considers four basic aspects of charts and graphs. The first of these is the general field or content area of the publication that contained the graph. We included six broad categories of content area: math, physical science, life science, social sciences, business and "General Interest", which is a catch-all category containing such items as magazines, newspapers and "How to" books. The second aspect considered by the sampling plan is the age level of the intended readership. This includes three mutually exclusive categories: pre-secondary, secondary, and adult. The third aspect considers the general format of the publication in which the graph appears. For adults we include journals, text books and general reading, whereas for younger readers, the source format is restricted to text books and general reading. The fourth and final aspect considered by the plan is the visual format of the display. The four categories of visual format are bar graphs, line graphs, pie graphs, and other graphs. Charts of any type were classified in the other graph category.

This sampling plan yielded 152 cells. Our initial intent was to find several graphs for each of these cells, however, after collecting over 300 graphs we were left with 77 empty cells. Most of these occur in the non-text book source format and in the nonadult age categories. After a heroic effort to fill the empty cells we concluded that charts and graphs were very infrequently used in these situations.

In using the sample for questionnaire development we first perused the entire sample and selected about twenty of the seemingly most problematical graphs. We then proceeded to isolate the particular operating principle violations in each graph, noting the graphic constitutents involved and the manner in which the violation occurred. These twenty "bad" graphs provided the foundation on which the original set of questions were framed. We continued to analyze more graphs selected from other cells of our sampling scheme to ensure wide applicability of the final questionnaire. In all, over ninety graphs were analyzed during the development of the questionnaire. As we discovered new ways in which operating principles were violated we added new questions or rephrased existing questions so that they would draw attention to violations if they exist. The questionnaire development was thus an iterative process of continually fine-tuning questions, making some more general, others more specific, until there was some assurance that most violations would be caught. There obviously is still room for improvement, however, as we demonstrate in the next section of this chapter, the current version is an excellent compromise in terms of problem areas detected for the amount of time and effort expended by the analyst.

Questionnaire Structure

The questionnaire is presented in the following section. It consists of two modules. The first and larger of these is concerned with single framework graphs or the individual subgraphs of a multiple framework display. The second module contains questions on the relationships between the various subgraphs of a multiple framework display. Each module is organized into three parts. The first contains questions pertaining to operating principles at the syntactic level. The second contains questions at the formal and semantic levels, and the third is concerned with pragmatic operating principles. Table 5.1 shows the number of questions in the questionnaire which pertain to each operating principle.

INSERT TABLE 5.1 HERE

Each of these divisions into levels of analysis is further organized by graphic constituent and combinations thereof. At the syntactic level each constitutent may be treated in isolation since syntax is concerned only with the processing of marks on the page. At the formal, semantic and pragmatic levels, however, the meanings and implications of constituents are determined by relations between them (i.e., how does the specifier operate in conjunction with the framework) as well as by the constituents themselves. Table 5.2 shows the number of questions in module 1 dealing with each constituents or combinations of constituents. Module 2 contains ten questions, all of which deal with the relationship between subgraphs in multiple framework displays.

INSERT TABLE 5.2 HERE

For the most part the structure described above is maintained throughout the questionnaire, however, there are one or two instances where a question concerning a principle at one level is placed among questions at a different level of analysis. An example of this is question 49 which concerns the syntactic principle of perceptual distortion. This question is included in the semantics section of the questionnaire because the reader is not likely to realize that this principle has been violated until he or she tries to assign meaning to the specifier. Perceptual distortion is truly a syntactic issue, however, since it is concerned with how the perceptual system processes a visual form.

In module 1 there are three spaces provided after each question. These are to be used to record independent responses for as many as three subgraphs in a multiple framework display. Of course only one space is needed for a single framework graph.

Each question in the questionnaire offers several alternative responses each of which falls into one of these categories. Some responses imply that the graph has no problem with regard to the issue addressed by the question. Other responses imply that the operating principle involved has been violated to a minor degree. These violations may be purely technical, causing no real impediment to the chart or graph, or they may result in a minor initial confusion on the part of the graph reader. Responses in this category are referred to as violations and are indicated by a single asterisk (*). Finally, the third category of responses corresponds to severe violations of an operating principle which either cause a great deal of confusion before being resolved or render some facet of the graph completely uninterpretable. Responses of this type are referred to as faults and are indicated on the questionnaire by a double asterisk (**).

A Tool for the Analysis of Charts and Graphs

The following questionnaire can be used with either charts or graphs. However, many of the questions are inappropriate for charts; these questions are preceded by the symbol, (@). In addition, because graphs are so much more frequently used, we have used some terminology that is specific to graphs, in particular we have referred to frameworks as being composed of axes. In a graph, the framework defines the domain and range that are related together by a specifier. This usually is an independent variable (things being varied, such as time) and a dependent variable (such as tons of wheat), with a line or bars serving as a function relating the two. In a chart, the framework is usually broken down into a set of boxes (as in a flowchart) or nodes (as in a family tree), and the specifier is composed of a set of lines that relate these framework elements together in some way (in a linear sequence, hierarchically, etc.). Thus, when analyzing a chart, simply substitute "box" or "node", as appropriate, when reading "axis". The same principles apply in the same way to both charts and graphs.

The reader should approach this questionnaire with a major caveat in mind: we can reveal violations of principles, but we cannot tell you whether these violations are important. That depends on the purposes to which the chart or graph is being put. Thus, after tallying up the sum of the violations, one must carefully consider the purpose of the display. Is it to express an idealization of some structure or relationship? If so, then great precision in internal mapping may not be required. Is it to display actual data precisely? If so, then virtually all violations will be important. Is it to lead the reader toward particular point of view? If so, then one may decide to violate the principle of invited inference. And so on.

The questions in each of the two modules (single display, multiple display) are numbered, and each bears a three letter code which associates the question with a specific level of analysis and operating principle. A list of the principles and codes is given on the next page.

Operating Principles and Associated Codes

Syntactic Principles

S-AD	adequate discriminaility
S-PL	perceptual distortion
S-GO	gestalt organization
S-DS	dimensional structure
S-PP	processing priorities
S-PL	processing limitations

Semantic Principles

M-SC	Surface Compatibility
M-SA	Schema Availability

Formal Principles

F-IN	internal mapping
F-EX	external mapping

Pragmatic Principles

P-II	invited inference
P-CX	context

Module 1: Analysis of a single display

Syntax

QUESTION#	PRINCIPLE		S[1]	S[2]	S[3]
		Outer Framework			
1	S-AD	Are the marks defining or implying the outer framework sufficiently discriminable such that the function of this constituent is recognizable? 1 yes 2 a brief search is required for recognition (*) 3 the function of the framework is very difficult to apprehend (**)	___	___	___
2	S-GO	Are Gestalt factors applied to the marks which define or imply the outer framework such that the correct organization is easily perceived? 1 yes 2 a brief search is required for recognition (*) 3 framework is very difficult to perceive (**)	___	___	___
@ 3	S-PL	If hash marks are used to subdivide intervals between scale value labels along the axes, is the number of marks small enough so that it can be apprehended at a glance? 0 not applicable 1 yes 2 no, some thought is required (*) 3 no, a great deal of thought is required (**)	___	___	___

QUESTION# PRINCIPLE S[1] S[2] S[3]

Inner Framework

If there is no inner framework, go to question 7.

@ 4 S-AD Are the marks defining or implying the inner framework sufficiently discriminable such that this constituent is recognizable?

0 not applicable
1 yes
2 a brief search is required for recognition (*)
3 framework is very difficult to perceive (**)

___ ___ ___

@ 5 S-GO Is the organization of marks defining or implying implying the inner framework sufficiently clear such that the constitutent is recognizable?

0 not applicable
1 yes
2 a brief search is required for recognition (*)
3 framework is very difficult to perceive (**)

___ ___ ___

@ 6 S-GO Is the relation between the inner and outer framework clear?

1 yes
2 yes, but some thought is required (*)
3 no (**)

___ ___ ___

7 S-PP Are the more visually salient features of either the inner or outer framework more important than the less salient features?

1 yes
2 no (*)

___ ___ ___

QUESTION#	PRINCIPLE		S[1]	S[2]	S[3]
		Background			
8	S-PP	If there is a background, are the background figures or designs too dominant such that they obscure the presentation of information? (If there is no background, go to question 10.) 1 no 2 yes, background figures hinder the presentation somewhat (*) 3 yes, background figures severely hinder the presentation (**)	___	___	___
9	S-PL	Does the number or complexity of background figures tax processing, leading to confusion? 1 no 2 yes (**)	___	___	___

QUES-TION#	PRIN-CIPLE		S[1]	S[2]	S[3]

Specifiers

10 S-AD Please apply the following categories to describe the case of discriminability of those visual continua, listed below, which are applicable to this chart or graph.

0 not applicable
1 levels or variations are easily discriminable
2 some difficulty in discriminating (*)
3 much difficulty in discriminating (**)

		S[1]	S[2]	S[3]
A	shape	___	___	___
B	length	___	___	___
C	size	___	___	___
D	position	___	___	___
E	orientation	___	___	___
F	continuity	___	___	___
G	lightness (achromatic)	___	___	___
H	lightness (chromatic)	___	___	___
I	hue	___	___	___
J	saturation	___	___	___
K	numerousness	___	___	___
L	other	___	___	___

11 S-GO Do the Gestalt principles (similarity, good form, symmetry, proximity and good continuation) imply conflicting organization of specifiers? (i.e., are any Gestalt factors displayed such that the correct organization is de-emphasized?)

1 no
2 yes, a slight conflict exists which can easily be resolved (*)
3 yes, a severe conflict exists which cannot easily be resolved (**)

___ ___ ___

QUESTION# PRINCIPLE S[1] S[2] S[3]

Specifiers (Cont'd)

@ 12 S-DS Is information concerning two distinct semantic variables being conveyed by two integral visual dimensions (e.g. oil production and population of a country conveyed respectively by the height and width of a rectangle)?

1 no
2 yes, but only a minor difficulty results (*)
3 yes, and it is very difficult to extract information (**)

___ ___ ___

@ 13 S-PL If parts of specifiers are separately identified (e.g. each bar has several distinct and abutting segments), is the reader required to compare parts in order to extract information from the graph?

0 not applicable
1 no
2 yes, however the information conveyed by these parts is of secondary importance (*)
3 yes, and the information encoded in these parts is of major importance (**)

___ ___ ___

14 S-PP If a specifier is used as a depiction, such that variations in size, orientation, etc. are important, are these variations emphasized so that they are encoded?

0 not applicable
1 yes, differences are very discriminable
2 no, differences are only moderately discriminable (*)
3 no, differences are not discriminable (**)

___ ___ ___

QUESTION# PRINCIPLE S[1] S[2] S[3]

Specifiers (Cont'd)

15 S-AD Is any local region of the graph area so densely packed that it is difficult to identify and interpret individual specifiers?

1 no
2 yes, some difficulty results from local crowding (*)
3 yes, a great deal of difficulty results from local crowding (**)

___ ___ ___

16 S-PL Is the whole graph area so densely packed with information that one cannot understand the information presented?

1 no
2 yes, some difficulty results from global crowding (*)
3 yes, a great deal of difficulty results from global crowding (**)

___ ___ ___

17 S-PP Do all visually salient features of the specifiers bear information that is pertinent to the intended message of the graph?

1 yes
2 no, however this results in only a minor processing difficulty (*)
3 no, and this results in a severe processing difficulty (**)

___ ___ ___

18 S-PP Is the visual dominance of the elements consistent with the points to be made? (e.g., are the more salient features more important?)

1 no inconsistency
2 minor inconsistency (*)
3 severe inconsistency (**)

___ ___ ___

QUES- PRIN-
TION# CIPLE S[1] S[2] S[3]

Labels

19 F-EX Is there a title?

1 yes
2 no (*)

___ ___ ___

If there is no title, go to question 22.

20 S-PP Is the title easily recognizable by virtue of its size and position?

1 yes
2 no, some search is required (*)
3 no, very difficult to recognize (**)

___ ___ ___

21 S-AD Is the title legible?

1 yes
2 no, some effort is required (*)
3 no, a great deal of effort is required (**)

___ ___ ___

22 S-PP Is the subtitle easily recognizable by virtue of their size and position? (If there is no subtitle, go to question 24.)

1 yes
2 no, some search is required (*)
3 no, very difficult to recognize (**)

___ ___ ___

23 S-AD Is the subtitle legible?

1 yes
2 no, some effort is required (*)
3 no, a great deal of effort is required (**)

___ ___ ___

QUESTION# PRINCIPLE S[1] S[2] S[3]

Labels (Cont'd)

24 S-PP Is the caption or legend easily recognizable? (If there is no caption or legend, go to question 26.)

1 yes
2 no, some search is required (*)
3 no, a great deal of effort is required (**)

___ ___ ___

25 S-AD Is the caption or legend legible?

1 yes
2 no, some effort is required (*)
3 no, a great deal of effort is required (**)

___ ___ ___

26 S-GO Do Gestalt principles seem to associate variable names with the appropriate axis? (If there is no variable name on axes, go to question 28.)

0 not applicable
1 yes
2 no, some attention is required (*)
3 no, a great deal of attention is required (**)

___ ___ ___

27 S-AD Are variable names legible?

0 not applicable
1 yes
2 no, some effort is required (*)
3 no, a great deal of effort is required (**)

___ ___ ___

QUES- TION# PRIN- CIPLE S[1] S[2] S[3]

Labels (Cont'd)

@ 28 S-GO Are scale values easily associated with corresponding tick marks along the appropriate axis? (If there are no scale values, go on to question 30.)

1 yes
2 no, some attention is required (*)
3 no, a great deal of attention is required (**)

___ ___ ___

@ 29 S-AD Are scale values legible?

1 yes
2 no, some effort is required (*)
3 no, a great deal of effort is required (**)

___ ___ ___

@ 30 S-GO Are the units easily associated with the appropriate scale values? (If unit of scale values are not marked, go to question 32.)

[V-4]

1 yes
2 no, some effort is required (*)
3 no, a great deal of effort is required (**)

___ ___ ___

@ 31 S-AD If units are marked, are the marks legible?

1 yes
2 no, some effort is required (*)
3 no, a great deal of effort is required (**)

___ ___ ___

QUES- TION# PRINCIPLE S[1] S[2] S[3]

Labels (Cont'd)

32 S-GO Are variable or level names readily associated with the appropriate specifiers?

0 not applicable
1 yes
2 no, some attention is required (*)
3 no, a great deal of attention is required (**)

___ ___ ___

33 S-AD Are these variable or level names associated with specifiers legible?

0 not applicable
1 yes
2 no, some effort is required (*)
3 no, a great deal of effort is required (**)

___ ___ ___

If no legend is used to label specifiers, go to question 35.

34 S-PL Are too many pairs of items present in the legend such that it is difficult to remember associations?

1 no
2 yes (*)

___ ___ ___

35 S-PP Is the visual dominance and form of label elements consistent with points to be made?

1 yes
2 no, minor inconsistency (*)
3 no, severe inconsistency (**)

___ ___ ___

QUESTION# PRINCIPLE S[1] S[2] S[3]

Framework

@ 36 F-EX Do the syntactic cues regarding denseness or differentiation for any frame elements contradict the semantic implications regarding these qualities?

1 there is agreement
2 there is a contradiction (*)

___ ___ ___

37 F-EX Is there only one apparent way of interpreting each frame element?

1 yes
2 no, but after deliberation only one way is plausible (*)
3 no, and there is no way of resolving which interpretation is correct (**)

___ ___ ___

38 F-EX Is every necessary part of the framework clearly implied or present?

1 yes
2 not clear (*)
3 no (**)

___ ___ ___

39 M-SC If the framework depicts, is it clearly representative of the class for which it stands?

0 not applicable
1 yes
2 no, the depiction is somewhat misleading (*)
3 no, the depiction is very misleading (**)

___ ___ ___

QUESTION#	PRINCIPLE		S[1]	S[2]	S[3]
		Framework (Cont'd.)			
40	M-SA	Is the form of the framework likely to be understood by the intended reader? 1 yes 2 no, probably not (*) 3 no, certainly not (**)	___	___	___
41	F-EX	Do all parts of the framework play a role in its function? 1 yes 2 no, but this does not distract or confuse (*) 3 no, and this is confusing (**)	___	___	___
42	M-SA	Are the variables associated with each frame element likely to be understood by the intended reader (e.g., derivatives, integrals, or other higher math concepts or technical terms)? 1 yes 2 no, probably not (*) 3 no, certainly not (**)	___	___	___
@ 43	F-EX	Are any axes discontinuous or non-uniform? 0 not applicable 1 no 2 yes, in an obvious way 3 yes, but not in an obvious way (*)	___	___	___

QUESTION#	PRINCIPLE		S[1]	S[2]	S[3

Background

44 F-EX If background figures are present, are they easily interpreted as such, or are they confused with specifier elements?

0 not applicable
1 yes
2 no, some confusion occurs (*)
3 no, a great deal of confusion occurs (**)

___ ___ ___

QUES- PRIN-
TION# CIPLE S[1] S[2] S[3]

Specifiers

45 M-SC If specifiers depict, is depiction representative of class for which it stands?

0 not applicable
1 yes
2 no, the depiction is somewhat misleading (*)
3 no, the depiction is very misleading (**) ___ ___ ___

46 M-SC If specifiers represent symbolically, are representations easily connected to their refferents?

0 not applicable
1 yes
2 no, the association is somewhat counter intuitive (*)
3 no, the association is very counter intuitive (**) ___ ___ ___

47 F-E. Is there only one apparent way of interpreting each specifier? (For instance, is it clear whether specifiers are contiguous or overlapping?)

1 yes
2 no, but after deliberation only one way is plausible (*)
[V-9]
3 no, and there is no way of resolving which interpretation is correct (**) ___ ___ ___

48 F-IN Is three dimensional perspective used in a way such that some specifiers (or parts thereof) are altered in shape or size?

1 no
2 yes, relative comparisons of specifiers are somewhat non veridical (*)
3 yes, relative comparison of specifiers is very non veridical and/or mappings are obscured (**) ___ ___ ___

QUES-TION# PRIN-CIPLE S[1] S[2] S[3]

Specifiers (Cont'd)

@ 49 S-PD Are the specifiers presented in a way that allows the reader to make subjective quantitative comparisons of elements based on visual inspection which are in accord with the actual quantitative relationships?

1 yes
2 no, subjective estimates are systematically off by a small amount (*)
3 no, subjective estimates are systematically off by a large amount (**)

___ ___ ___

50 F-EX Are symbols representing different items differentiable?

0 not applicable
1 yes
2 no, some attention is required (*)
3 no, it is very difficult to apprehend differences (**)

___ ___ ___

51 F-EX Is every meaningful difference indicated clearly by a difference in marks?

1 yes
2 no (**)

___ ___ ___

52 M-SC Are the visual continua along which specifiers vary compatible with information displayed?

1 yes
2 no, slightly incompatible (*)
3 no, completely incompatible (**)

___ ___ ___

QUESTION#	PRINCIPLE		S[1]	S[2]	S[3]

Specifiers (Cont'd)

53 M-SC Is the spontaneous interpretation of the specifier compatible with the cognitive construct being represented?

1 yes
2 no, slightly incompatible (*)
3 no, completely incompatible (**)

___ ___ ___

QUESTION# PRINCIPLE S[1] S[2] S[3]

Labels

If any of the following types of labels are absent or incomplete, assess the impact on the interpretability of the graph. (In each case, please use one of the following responses.)

0 not applicable
1 easily interpretable
2 eventually interpretable (*)
3 uninterpretable (**)

54 F-EX Variable label on axes. ___ ___ ___

@ 55 F-EX Scale values on axes. ___ ___ ___

@ 56 F-EX Units of scale values. ___ ___ ___

57 F-EX Labels on specifiers. ___ ___ ___

58 M-SC If depictions serve as labels, are they clearly representative of the class of objects for which they stand?

0 not applicable
1 yes
2 no, the depiction is somewhat misleading (*)
3 no, the depiction is very misleading (**)

___ ___ ___

59 M-SA Are the words used in labels clear and comprehensible to the intended reader?

1 yes
2 no, probably not (*)
3 no, certainly not (**)

___ ___ ___

60 M-SA Are symbols used in labels familiar to the intended reader?

0 not applicable
1 yes
2 no, probably not (*)
3 no, certainly not (**)

___ ___ ___

QUES-TION#	PRIN-CIPLE		S[1]	S[2]	S[3]

Labels (Cont'd.)

61 F-IN If specifiers are labelled in a legend or caption, is the correspondence between items on the graph and those in the legend clear?

0 not applicable
1 yes
2 minor inconsistency (*)
3 severe inconsistency (**)

___ ___ ___

QUES-TION# PRIN-CIPLE S[1] S[2] S[3]

Framework x Specifiers

62 F-IN Is it clear which points of which axes are being related [illegible] each specifier?

0 not applicable
1 yes
2 no, it is not obvious immediately but can eventually be resolved (*)
3 no, and the correct correspondence cannot be determined (**)

___ ___ ___

@ 63 F-IN If any specifiers extend outside the region bounded by the frame elements, is there a consequent loss of precision in assigning quantitative values?

0 not applicable
1 no
2 yes, but the loss of precision is not important for most purposes (*)
3 yes, and the loss of precision severly hampers the use of the graph for its intended purpose (**)

___ ___ ___

@ 64 F-IN Is the level of precision of the scale markings and hash marks along the frame compatible with the mapping precision afforded by the specifiers?

1 yes
2 no (*)

___ ___ ___

QUES- TION# PRIN- CIPLE S[1] S[2] S[3]

Framework x Specifiers x Labels

65 F-EX Is there an apparent and logically consistent literal interpretation for each visually salient feature in the graph?

1 yes
2 no, minor confusion results (*)
3 no, a great deal of confusion results (**)

___ ___ ___

66 M-SC Are the various graphic elements and visual properties used in a way that is consistent with cultural conventions? (e.g., red implies danger, green implies safe. If symbols have an accepted meaning is their use consistent with this meaning? [(✓) implies okay, (x) implies incorrect])

0 not applicable
1 yes
2 no, cultural conventions have been ignored but not violated (*)
3 no, the use of graphic elements and visual properties is blatantly inconsistent with cultural conventions (**)

___ ___ ___

Pragmatics

QUESTION#	PRINCIPLE		S[1] S[2] S[3]

Framework

@ 67 P-II Are scale units, aspect, or use of truncation of axes proper for the impression the illustrator wishes to convey?

1 yes
2 no (**)

___ ___ ___

68 P-II If the framework depicts, does it convey a message consistent with the point of the graph?

0 not applicable
1 yes
2 no, a slight contradiction is implied (*)
3 no, a severe contradiction is implied (**)

___ ___ ___

QUESTION#	PRINCIPLE		S[1] S[2] S[3]

Background

69 P-II Do background figures, if present, convey a message consistent with the point of the graph?

0 not applicable
1 yes
2 no, a slight contradiction is implied (*)
3 no, a severe contradiction is implied (**)

___ ___ ___

QUES- TION# PRIN- CIPLE S[1] S[2] S[3]

Specifiers

70 P-II Do the specifiers make or imply inferences that are contradictory to messages conveyed elsewhere in the chart or graph?

1 no
2 yes, slight contradiction (*)
3 yes, glaring contradiction (**)

___ ___ ___

QUESTION# PRINCIPLE S[1] S[2] S[3]

Labels

71 P-CX Is the graph suitably introduced by:

1 title or subtitles
2 comments
3 caption
4 no, it is not suitably introduced (**)

(note: more than one answer is possible) ___ ___ ___

72 P-II Do connotations of labels agree with the visual impact of the display?

1 yes
2 no (*)

___ ___ ___

73 P-CX Are labels consistent with the terminology used in the text?

0 not applicable
1 yes
2 no (*)

___ ___ ___

QUESTION# PRINCIPLE S[1] S[2] S[3]

Framework x Specifiers x Labels

74 P-CX Is the information presented in the chart or graph compatible with the adjacent text?

0 not applicable
1 yes
2 no, the two are slightly inconsistent (*)
3 no, the two are very inconsistent (**)

___ ___ ___

75 P-CX Are the invited inferences compatible with the information presented in the adjacent text?

0 not applicable
1 yes
2 no (*)

___ ___ ___

Module 2: Organization of Subgraphs

Syntax

QUESTION#	PRINCIPLE		S[1]	S[2]	S[3]
76	S-PP	Does the physical arrangement of subgraphs lead the reader to examine them in a logical sequence? 1 yes 2 no, the arrangement is suboptimal (*) 3 no, the arrangement is very confusing (**)			
77	S-PP	Does relative visual saliency of subgraphs correspond to the relative importance of the information presented in each display? 1 yes 2 no, but only a moderate problem (*) 3 no, leading to confusion (**)			
78	S-PL	Are there too many subgraphs to comprehend at once? 1 no 2 yes (*)			
79	S-GO	Do the Gestalt factors lead one to make the appropriate associations between items in a legend and their referrents in the various subgraphs? 0 not applicable 1 yes 2 no, some confusion exists (*) 3 no, a great deal of confusion exists (**)			

Semantics

QUES- PRIN-
TION# CIPLE S[1] S[2] S[3]

80 F-IN If there are instances where the same variable is discussed in several subgraphs, is the physical arrangement conducive to comparisons at equal values for all shared variables?

0 not applicable
1 yes
2 no, however, in the context of the presentation such comparisons are not relevant
3 no, and the usefulness of the graph is slightly impaired (*)
4 no, and the usefulness of the graph is severely impaired (**)

___ ___ ___

@ 81 F-IN If there are instances where the same variable is discussed in several subgraphs, are the axis scales for the shared variables marked in the same units per inch?

0 not applicable
1 yes
2 no, however, this is in the best interest of communicating the information
3 no, and the usefulness of the graph is slightly impaired (*)
4 no, and the usefulness of the graph is severely impaired (**)

___ ___ ___

82 F-IN If one subgraph presents a second view of the information in another subgraph, is the correspondence between the two subgraphs clear?

0 not applicable
1 yes
2 no, some inspection is required (*)
3 no correspondence is evident (**)

___ ___ ___

QUESTION#	PRINCIPLE		S[1]	S[2]	S[3]
83	F-EX	Are the title, comments, and other labels, in conjunction with the graphic material, sufficient to explain the relationship between the various subgraphs? 1 yes 2 no, some aspects remain unclear (*) 3 no, the overall relationship between subgraphs remains unclear (**)	___	___	___
84	F-EX	If lines or other marks are used to relate subgraphs, is it clear how each of them functions? 0 not applicable 1 yes 2 some doubt exists (*) 3 probably not (**)	___	___	___
85	F-IN	If the different subgraphs employ constituents of different forms serving the same purpose, does this increase the workload on the reader? 0 not applicable 1 no 2 moderately (*) 3 yes (**)	___	___	___

Questionnaire Reliability

The success of our questionnaire as an evaluative aid hinges on its reliability. Reliability, as the term is used here, concerns the degree to which different analysts using the questionnaire to analyze a given chart or graph agree in their analysis. Obviously the questionnaire would be worthless if each independent person using it to evaluate the same graph produced a different set of violations.

The way in which we assessed the reliability of the questionnaire is best illustrated by considering the possible outcomes of an analysis of a chart or graph by two different analysts. The possible outcomes are presented in Table 5.3.

INSERT TABLE 5.3 HERE

One outcome has analyst 1 scoring a problem with respect to a particular operating principle while analyst 2 scores no such problem. A second outcome has both analysts agreeing that no problem occurred. A third outcome has analyst 1 scoring no problem while analyst 2 scores a problem. Finally, the fourth outcome has both analysts agreeing that a problem has occurred. If we divide the total number of agreements (a+d) by the total number of questions on which agreements were possible (a+b+c+d), we get an agreement rate, r,

$$r\% = \frac{a+d}{a+b+c+d} \times 100 \qquad (1)$$

Note that this rate can vary from 0% (perfect disagreement) to 100% (perfect agreement).

The overall agreement rate for our questionnaire was determined by aggregating the possible outcomes of two analysts who independently evaluated ten separate graphs (four of which were multiple framework) randomly selected from the sample. Both violations (*) and faults (**) were classed as equivalent and formed a single category called problems. This category was then contrasted with the category called no problems. One analyst was very experienced with the scheme and the other analyst was naive at the outset. Because there are 75 questions in module 1 (dealing with single framework graphs) and 10 additional questions in module 2 (dealing with multiple framework graphs) there is a total of 790 questions from which agreement and disagreement is based. The actual values are presented in Table 5.4. From this table, the rate is 96.58 percent indicating a fairly strong agreement between the two analysts on what constituted a problem in this group of graphs.

INSERT TABLE 5.4 HERE

A closer examination of the data reveals a greater than 97 percent agreement rate between the analysts on seven of the ten graphs. Of the remaining graphs, only one fell below the 90 percent rate, a low of 88 percent. An examination of the data associated with this worst case reveals more than half the disagreements between the analysts concerned the formal principle of external mapping. This is not surprising since overall, ignoring the particular violation cited, and simply noting whether there is agreement about a syntactic, semantic, formal or pragmatic violation, the formal analysis yields the lowest agreement rate, a 94.39 percent rate. This is to be contrasted with the 95.45 percent rate for the closely aligned semantic principle, and 97.00 percent and 98.09 percent agreement rates for the pragmatic and syntactic principles respectively. Similarly, if we simply rate whether the analyst spotted a problem with the framework, specifier, labels or background there was high agreement on localizing problems to these basic constitutents with rates ranging from a low of 95.78 percent for the specifier to a high of 97.65 percent for the framework.

The upshot of these analyses is straightforward; even a naive analyst is able to use the questionnaire to reveal basic problems with a graphic display and the instrument is quite reliable. It is worth noting in addition that the analysts almost always saw the same problems in a display, but differed in the way the conceptualized (and hence categorized) the problems. The bases for such differences will be developed in the following chapter.

Application of the Questionnaire to a Representative Sample of Charts and Graphs

In the previous sections we have shown the questionnaire to be reliable in that two analysts independently discovered essentially the same violations in a set of ten graphs. In this section we report the results of applying the questionnaire to a substantial subset of the charts and graphs which were collected according to our sampling scheme. The purpose of this effort was two-fold: First it served as a final test of the questionnaire on a diverse set of graphs taken from commonly encountered reading material, and second, it provides a description of patterns of operating principle violations in various categories of charts and graphs.

The sampling scheme was described in detail in a previous section. We randomly selected one graph from each of the 75 non-empty cells of the sampling scheme (see Figure 5.1) and divided these between two analysts who then applied the questionnaire to each graph.

The reader can get a detailed view of how the non-empty cells of the sampling scheme are distributed with respect to chart and graph categories in Figure 5.1, however, Table 5.5 summarizes the gross features of the distribution. Note from the table that the numbers of graphs analyzed were fairly evenly distributed amongst the content areas. Most of these graphs, though, were found in adult journals and textbooks indicating a dearth of this material in publications for children. Similarly, the distribution of formats correspond to the predominance of bar charts and line graphs found in the literature.

INSERT TABLE 5.5 HERE

The distribution of the number of faults or serious violations per graph for each of these categories is summarized in Table 5.6. From this table, the distribution of faults within a category follows the independence model for all categories with the exception of content area. For this category, graphs found in the business area result in a much higher number of faults than graphs found in other areas. This observation is confirmed by a reliable chi-square, $\chi^2(5) = 13.56, p<.02$. A closer examination of the data for this case reveals that these graphs are not as well executed as graphs found in the other content areas. That is 41.1 percent of all faults concerning principles pertinent to organization (e.g., similarity, proximity, etc.) and 38.9 percent of all faults concerning principles pertinent to seeing the lines (e.g., discriminability) occurred in this content area. Additionally the graphs sampled from this content area employ symbolic representations that are not as compatible with their reference nor as consistent with the conventions of our culture as graphs found in other areas. This is evidenced by the fact that of all the faults concerning the principle of surface compatibility, 62.5 percent were found in graphs taken from the business area.

INSERT TABLE 5.6 HERE

The distribution of faults per question set as a function of the different levels of description and the distribution of faults per question set for each of the different operating principles are shown in Tables 5.7a and 5.7b respectively. While neither table is consistent with the independence hypothesis, some comments are in order. Note that faults pertaining to formal principles, specifically the principle of external mapping occur most frequently. Recall that this principle pertains to the meaning given the marks and is violated if a mark is ambiguous or a necessary set of marks is missing.

The discussion in Chapter 4 lists some of the pitfalls to be avoided. Additionally, the use of a title as a descriptive aid will help orient the reader and thus reduce potential ambiguity.

INSERT TABLE 5.7 HERE

Also note from Table 5.7b that another area in need of improvement involves the organization of the marks. Most often carelessness is the culprit in this case. A careful analysis of the graph upon completion should help reduce violations of this principle.

Table 5.8 shows the distribution of faults per question set as a function of the different graphic constitutents and their combinations. The table shows the greatest proportion of faults pertains to the specifier alone and its interaction with the framework. The incidence of faults for this basic level constituent is roughly two times greater than the other basic levels and this increased incidence is reliable, $\chi^2(6)=21.19$, $p<.01$.

INSERT TABLE 5.8 HERE

The breakdown of the proportion of faults, in terms of specific operating principles violated, for these two constituents is shown in Table 5.9. From this table, the proportion of faults pertaining to the interaction between framework and specifier is entirely due to violations of the internal mapping principle. Violations, in this case, usually result when the graph maker tries to represent two or more range scales on a single framework graph, having multiple specifiers. If such a condition is envisaged it may be preferable to represent the information using multiple frameworks. With regard to the specifier, we note, as before, that most of the proportion of faults is accounted for by violations of the external mapping and discriminability princples.

INSERT TABLE 5.9 HERE

In summary, the pattern of results reported in this chapter suggest two areas of improvement in graph design. First, we must ensure that the interpretation of the graph is not contaminated by either the addition of too much information (such as the careless placement of a second range scale on a single framework graph) or the deletion of relevant information (such as a title). Second, we should exercise care in the execution of the graph by ensuring that adequate discriminability and organization result over a wide range of conditions and graph readers.

Two Graphs Seen in a New Light

At the beginning of Chapter 1 we saw two graphs (Figures 1.1 and 1.2) that were obviously flawed. However, most people could not say exactly how these displays were amiss, but had only haphazard intuitions and sketchy diagnoses. Let us return to those graphs now, armed with the diagnostic tool just presented.

Figure 1.1: Falling interest rates.

In using the questionnaire the following violations of Figure 1.1 were revealed:

Syntax:

The specifier violated two syntactic principles. The principle of adequate discriminability was violated because the colors of the lines were too similar, making it almost impossible to tell them apart when they crossed. The principle of processing priorities was also violated because the specifier lines differed dramatically in their saliency, but this difference did not reflect any difference in the importance of the information being presented.

The labels violated one syntactic principle in three different places. First, the principle of gestalt organization was violated because scale values were not associated with the tick marks; second, it was violated because units were not associated with the scale values; and third, it was violated because variable or level names were not associated well with the specifiers (at the far right).

Semantics:

The framework violated one semantic principle, that of representativeness. The framework did not clearly depict a bank, and this depication was, therefore, slightly distracting.

Formal:

The framework violated the external mapping principle because its markings were not consistent with the concept being represented; although this is a minor point in the present case, the graph would have been useful more generally if the axes had been marked into discrete units.

The specifier violated the internal mapping principle because the foreshortening of the framework resulted in difficulty in comparing the form of the functions at different places in the graph. That is, the slopes must be mentally adjusted to compensate for the distorted framework in order to compare slopes at different points along the specifiers.

The graph as a whole violates the external mapping principle because there is no logically-consistent interpretation for all of the marks; for example, why are some labels along the axes inside the framework whereas others are outside the framework?

Pragmatics:

Finally, the pragmatic principle of invited inference was violated because of the foreshortening of the framework: although this does succeed in making the fall seem steeper, it also makes the rise seem steeper--which is an accidental byproduct of the attempt at distortion.

Figure 1.2: Ecological niches

This graph was baffling to many people; our system explains why.

Syntax:

The specifiers violated four of our syntactic principles: First, the princi[illegible] of dimensional structure was violated because peopl[illegible] see rectangles--and the specifier is in fact two distinct extents (one horizontal and one vertical); because the dimensions of sides of rectangles are integral, the value along one extent cannot help but influence how we see the value along the other. Second, the principle of adequate discriminability is violated because some of the specifiers are hard to see. Third, the principle of processing priorities is violated because the most visually striking specifiers are not necessarily the most important. Fourth, the principle of gestalt organization is violated because when two of the specifiers overlap, new rectangles are formed by the patch of common color. But these rectangles do not represent additional specifiers, and hence are very misleading.

The lables violated two syntactic principles. The title is difficult to see, violating the principle of adequate discriminability. The keys are divided into two segments, and group via proximity to distinct panels--even though all six key elements are relevant to each panel. Thus, the key violates the principle of gestalt organization.

Formal:

Finally, this graph violates a formal principle. The external mapping principle is violated because there is more than one way of interpreting the specifier elements.

Thus, it is clear that our analytic system not only provides insight into what previously was pretty murky territory, but generalizes to displays quite unlike those that originally shaped it (at the end of Chapter 2). The analytic scheme has now gone about as far as possible given the level of sophistication of the psychological theorizing engaged in up until now. Thus, in the next chapter we will consider further developments in the context of developing a detailed theory of graph reading per se.

CHAPTER 6: A THEORY OF GRAPH COMPREHENSION

So far we have concentrated on analyzing existing displays. But it is far better to draw a good graph to begin with than to correct one after the fact. In order to know how to generate an effective graph, however, we must have a theory of how a reader will actually process the display while reading it. This theory can then be used to guide the graph maker to construct an unambiguous, effective display. The simplified treatment of visual information processing presented in Chapter 2 is too sketchy to serve these ends. Thus, in this chapter we will consider in detail how people come to understand charts and graphs.

I. Introduction

Unlike seeing in depth, uttering a sentence, or reaching for a target, comprehending a graph is not something that anyone could argue is accomplished by a special-purpose mental faculty. Graphs are a recent invention in the history of our species, and if they are an especially effective method of communication, it must be because they exploit general cognitive and perceptual mechanisms in an optimal way. A theory that hopes to explain the process of graph comprehension will have to identify the psychological mechanisms used in interpreting a graph, and a theory that hopes to lead the way to more comprehensible graphs and more efficient graph readers will have to specify which operating principles of each mechanism contribute to the overall ease or difficulty of a graph. Thus, a theory of graph comprehension will draw heavily on general cognitive and perceptual theory, and where our knowledge of cognitive and perceptual mechanisms is sketchy, we can expect corresponding gaps in our ability to explain the understanding of graphs. The worth of a theory will probably lie not so much in its current successes in accounting for data and guiding the graph maker as in its promise of offering deeper and deeper explanations of graph comprehension as it absorbs the future discoveries of cognitive science.

As was revealed in our survey, there is a bewildering variety of graphs in current use, ranging from the line and bar graphs common in scientific journals, to drawings in popular magazines in which the thicknesses of two boxer's arms might represent the missile strength of the US and USSR, or in which the lengths of the rays of light emanating from a yellow disc might represent the price of gold in different months. Nonetheless, all graphs can be given a common characterization. Each graph tries to communicate to the reader a set of pairings of values on two or more mathematical scales, using objects whose visual dimensions (i.e., length, position, lightness, shape, etc.) correspond to the respective mathematical scales, and whose values on each dimension (i.e., an object's particular length, position, and so on) correlate with the values on the corresponding scales. The pairing is accomplished by virtue of the fact that any seen object can be described simultaneously by its values along a number of visual dimensions. For example, Figure 6.1 represents a pairing of values on a nominal scale (countries) with values on a ratio scale (GNP) using objects (bars) whose horizontal position (a visual dimension) corresponds to a value on the first scale, and whose height (another visual dimension) corresponds to a value on the second scale.

INSERT FIGURES 6.1 AND 6.2 HERE

Figure 6.2 represents a pairing of values on an ordinal scale (months) with values on an interval scale (temperature) using objects (wedges) whose radial position represents the month, and whose darkness represents the temperature. This characterization, which can be applied to every graph we have seen, was first pointed out by Bertin (1967) in his seminal treatment of charts, graphs, and maps.

As Bertin points out, this characterization implies that a graph reader must do three things: a) identify, via alphanumeric labels, the conceptual or real-world referents that the graph is conveying information about (Bertin calls this "external identification"), b) identify the relevant dimensions of variation in the graph's pictorial content, and determine which visual dimensions corresponds to which conceptual variable or scale (Bertin's "internal identification"), and c) use the particular levels of each visual dimension to draw conclusions about the particular levels of each conceptual scale (Bertin's "perception of correspondence").

Even a characterization as simple as this one raises a host of psychological questions, and until these questions are answered, we will not be able to predict what will make a particular graph easy or difficult to comprehend. These questions subdivide into two classes. First, note that a graph reader must mentally represent the objects in the graph in only a certain way. In the case of Figure 6.1, he or she must think of the bars in terms of their positions on the page, the jagged contour formed by the tops of the bars, their left-to-right order, and so on. This raises questions about how a visual stimulus is encoded internally, or, in the terms of the theory we will outline here, how the reader's visual description of the graph is built up. Second, the graph reader must remember or deduce which aspects of the visual constituents of the graph stand for which of the mathematical scales that the graph is trying to communicate. This raises questions about how knowledge in memory interfaces with visual input, or, in the terms of the present theory, how the reader's graph schema will spell out the ways in which the physical dimensions of the graph may be mapped onto the appropriate mathematical scales. In using the "visual description" and the "graph schema" to interpret a graph, a reader may obtain different sorts of information about it. Bertin points out that a reader can extract the exact value of some scale paired with a given value on another scale, the rate of change of values on one scale within a range of values on another, a difference between the scale values of two entities, and so on. We will use the term conceptual question to refer to the particular sort of information that a reader wishes to extract from a graph, and conceptual message to refer to the information that the reader, in fact, takes away from it (cf. Bertin, 1967).

In the rest of the chapter, we go beyond Bertin's work by defining and characterizing each of the mental representations involved in graph comprehension, proposing ways in which they are constructed and transformed in the course of reading a graph, and attempting to outline principles that dictate which aspects of these processes and representations affect the ease of extracting a message from a graph. These principles will provide the theoretical basis for the operating principles discussed in the preceding chapters, replacing the simplified theory of visual information processing presented in Chapter 2. We will try to justify these proposals by appealing to existing knowledge of perceptual and cognitive functioning, and by showing concrete instances of graphs and other visual displays whose degree of intuitive difficulty is explained by the proposals. Of course, the ultimate empirical test of the theory will be its ability to explain the relative ease with which various sorts of people extract various sorts of information from various sorts of graphs, over as wide a range of people, messages, and graphs as possible.

II. The Visual Array

The information in a graph arrives at the nervous system as a two-dimensional pattern of intensities on the retinas. We will use the term visual array to refer loosely to those early visual representations that depict the input in a relatively unprocessed, pictorial format (cf. the "2-1/2 dimensional sketch" of Marr & Nishihara, 1977, and the "surface array" of Kosslyn, Pinker, Smith, & Shwartz, 1979). Information in this form is, of course, far too raw to serve as a basis for comprehending the meaning of the graph. For that, we need a representational format that can interface easily with the memory representations embodying knowledge of what the visual marks of the graph signify. Such memory representations cannot be stated in terms of specific distributions of light and dark as would be represented in the visual array, because vastly different intensity distributions (differing in size, orientation, color, shape, lightness, etc.) could all be equivalent exemplars of a given type of graph. Thus, the representation that makes contact with stored knowledge of graphs must be more abstract than a visual array.

III. The Visual Description

A fundamental insight into visual cognition is that the output of the mechanisms of visual perception is a symbolic representation or "structural description" of the scene, specifying the identity of its parts and the relations among them (see Winston, 1975; Marr & Nishihara, 1977; Palmer, 1975; Pylyshyn, 1973). This mental description is not in English, of course, but in some symbolic "language of thought" which represents visual information in a manner appropriate to its use by other cognitive processes such as language, reasoning, motor control, and so on. In this description, the various aspects of the scene, such as its constituent elements, and their size, shape, location, color, texture, etc., together with the spatial relations among them, will be factored apart into separate symbols. As a result, each higher-level cognitive process need only refer to the symbols representing the aspect of the scene that is relevent to its own computations. For example, processes governing limb movement will access symbols explicitly representing an object's position in the three-dimensional world, whereas processes that formulate the sequence of words that will be uttered in response to the question "What color is that shirt?" will access symbols explicitly representing an object's hue.

This allows us to describe the mind economically as a set of more-or-less autonomous modules (see Simon, 1969): there is a visual system which need "know" nothing about either English syntax or skeletal musculature and, a linguistic system, which need "know" nothing about the laws of perspective, and a motor control system which need "know" nothing about the laws of color mixture--all the systems can communicate via a common symbolic description of a scene. We will use the term visual description to refer to the structural description representing a graph, and visual encoding processes to refer to the mechanisms that create a visual description from a visual array pattern.[1]

Many "languages" for visual descriptions have been proposed in the literature on vision in psychology and artificial intelligence (e.g., Palmer, 1975; Marr & Nishihara, 1977; Hinton, 1979; Winston, 1975; Miller & Johnson-Laird, 1976). Most of them describe a scene using propositions, whose variables stand for perceived entities or objects, and in which predicates specify attributes of and relations among the entities. It is assumed that the visual encoding mechanisms can detect the presence of each of these predicates in the visual array. For example, one-place predicates specify a simple property of an object, such as Circle (x) (i.e. "x is a circle"), Convex (x), Curve (x), Flat (x), Horizontal (x), Linear (x), Small (x), and so on. Two-place predicates specify the relations between two objects, such as Above (x,y) (i.e. "x is above y"), Adjacent (x,y), Below (x,y), Higher (x,y), Included-in (x,y), Points-towards (x,y), Parallel (x,y), Part (x,y), Near (x,y), Similar (x,y), Top (x,y), and so on. Three and higher-place predicates indicate relations among groups of objects, such as Between (x,y,z) (i.e., " x is between y and z"), In-line (x,y,z), and so on. Parameterized predicates take a number of variables and a number of quantitative constants, such as Area (x,α) (i.e., "x has area α"), Width (x,α), Location (x,α,β), Lightness (x,α), Orientation (x,α), and so on. These predicates may also be appropriate for specifying continuous multidimensional attributes of objects, which otherwise would be difficult to specify by a predicate chosen from a finite list. For example, any member of a class of shapes ranging from a flattened horizontal ellipse through a circle to a flattened vertical ellipse can be specified by two parameters, representing the lengths of the major and minor axes of the ellipse, thus: Ellipse (x,α,β).

---

[1] Note that our use of structural descriptions to represent the information in a graph does not bear on the devate over whether mental images involve information in an array or a structural description (e.g., Kosslyn, Pinker, Smith and Schwartz, 1979). That debate is not over whether arrays and structural descriptions exist in general, but whether the array can be filled with information from long term memory as well as from the eyes.

As is fitting for a paper on graphs, we will use a graphic notation for visual descriptions. Each variable in a description will be represented by a small circle or node in which the variable name is inscribed (for simplicity's sake, we will usually omit the variable name in the diagrams to be used in the chapter); each one-place predicate will simply be printed next to the nodes representing the variables that they are true of; and each two-place predicate will be printed alongside an arrow linking the two nodes representing the predicate's two arguments. Thus, a particular scene represented as the visual array in Figure 6.3a will be represented as the visual description in Figure 6.3b, or its graphic counterpart in 6.3c.

INSERT FIGURE 6.3 HERE

Constraining the Visual Description

If, as we argued, a visual array representation is unsuitable for the computations involved in extracting information from a graph, an unconstrained visual description is not much better. Since any visual array can be described in an infinite number of ways, a theory that allowed any visual description to be built from a visual array would be unable to predict what would happen when a given individual faced a given graph. For example, the array in Figure 6.3a can give rise not only to the visual description in Figure 6.3c, but to the descriptions in Figure 6.4 as well.

INSERT FIGURE 6.4 HERE

Clearly, if it is not to be utterly vacuous, the theory must specify which visual description is likely to be constructed in a given situation, based on our knowledge of how the human visual system works. In the following section, we summarize four broad principles, each grounded in basic psychological research, which constrain the form of visual descriptions. These principles will bear a large explanatory burden in the theory to be outlined here, since later we will claim that a prime determinant of the difficulty of a graph will be whether the visual description specifies explicitly the visual dimensions and groupings that the graph maker recruited to symbolize the mathematical scales involved in the message of the graph.

A. The Indispensibility of Space

It has long been known that an object's spatial location has a different perceptual status than its color, lightness, texture, or shape. Bertin (1967) tries to formulate this generalization by distinguishing between the two spatial dimensions of the surface of the paper (his "dimensions du plan", loosely, "framework dimensions") and other dimensions such as lightness and color (which he calls "retinal dimensions"). Michael Kubovy (1981) has addressed this issue systematically, and calls the two spatial dimensions of vision (plus the time dimension) indispensible attributes, analogous to the dimensions of pitch and time in audition. He defines the term "indispensible attribute" as an attribute with the following properties:

1) Perceptual Numerousity. The first constraint on a visual description must be on what is to count as a variable or node. Variables should stand for perceptual units of some sort, and not for any arbitrary subset of the light reflected from a scene (e.g., the set of all light patches whose dominant wave-

length is divisible by 100).[2] Kubovy points out that our perceptual systems pick out a "unit" or an "object" in a visual scene as any set of light patches that share the same spatial position, but not as a set of light patches that share some other attribute such as wavelength, intensity, or texture. Thus, Figure 6.5a will give rise to the visual description in Figure 6.5b, which partitions the array into three variables according to spatial location, rather than that in Figure 6.5c, which partitions the array into two variables according to surface markings.

INSERT FIGURE 6.5 HERE

2) Configural Properties. The second constraint on a visual scene is the choice of predicates available in assembling a visual description. Naturally, there will be predicates corresponding to all perceptible physical dimensions (e.g., bright (x), red (x), shiny (x), lightness (x, α), length (x, α); in addition, there will be "configural" or "pattern" predicates corresponding to higher-order functions defined over the physical dimensions. Kubovy points out that most configural properties in a sensory modality are defined over the indispensible attributes, which in the case of static objects vision are the vertical and horizontal spatial dimensions. As a consequence, there exist many predicates for spatial shapes (each of which can be defined by certain well-developed changes in relative horizontal and relative vertical positions in a pattern), but few for nonspatial "shapes" defined by analogous well-defined changes in other dimensions. For example, the array in Figure 6.6a contains elements whose heights increase with their horizontal position (lightness vary-

---

[2] This question, incidentally, is begged by Bertin's proposal that the difficulty of a graph may be predicted by how many "perceptual glances" a reader must make in reading a graph. Until we know what forms a perceptual unit that a "glance" centers upon, we will not know how many glances must be made.

ing randomly); the array in Figure 6.6b contains elements whose lightnesses increase with their orientations (position varying randomly).

INSERT FIGURE 6.6 HERE

However, the increase is immediately noticeable only in Figure 6.6a, where the increase is of one spatial dimension with respect to another, not in 6.6b. Correspondingly, there exists a predicate diagonal (x) that can be used to describe the scene in 6.6a, but nothing analogous for describing the scene in 6.6b, whose elements would probably be specified individually. Note that as long as one member of a pair of related dimensions is spatial, there may be configural predicates available; when neither member is spatial, configural predicates are unlikely. Thus, the elements in Figure 6.7 get darker with height, a change that, unlike that in 6.6c, is quickly noticeable, and may be captured by a single predicate [e.g., darkens (x)].

(Insert Figure 6.7 Here)

3) Discriminability and Linearity. As we review in Chapter 3 of this volume, physical variables are not in general preceived linearly, nor are small differences between values of a physical variable always noticed. In the visual description, this corresponds to numerical variables (e.g., height (x,17)) being distorted with respect to the real world entities they represent, or to distinct numerical variables sharing the same value when the represented entities in fact differ (e.g., lightness (x,17); lightness (y,17) for two boxes differing slightly in lightness). Kubovy remarks that indispensible attributes afford finer discriminations and more linear mappings than dispensible attributes, and indeed, our summaries in Chapter 3 show that the Weber fraction for spatial extent is 0.04, and the Stevens exponent is 1.0, both indicating greater accuracy for the representation of spatial extent than for the representation of other physical variables.

4) Selective Attention. As a consequence of (1), each variable may have associated with it a unique pair of coordinates representing its location. This means that location could serve as an index or accessing system for visual information. This is a form of selective attention, and Kubovy summarizes evidence supporting the hypothesis that attention is more selective for the indispensible attributes (horizontal and vertical location) than for other visual attributes (e.g., one cannot easily attend to all visible objects with the same lightness or shape, regardless of location, see Posner and Snyder, 1980, for example). In the theory outlined in this chapter, selective attention according to location will consist of a mechanism that activates various encoding mechanisms to process a given spatial region of the visual array, in order to encode more predicates into the visual description or to verify whether a given predicate is true of the entity at that location.[3] As we shall see, these mechanisms will play an important role in the "question-driven" or "top-down" processing of graphs.

B. Gestalt Laws of Grouping

The principles associated with the indispensibility of space in vision place constraints on the portions of an array that variables may stand for, on how numerical variables represent physical continua, and on how predicates are encoded or verified with respect to the visual array. What is needed in addition is a set of principles governing how variables representing visual entities will be related to one another in visual descriptions, that is, how the atomic perceptual units will be integrated into a coherent percept. A notable set of such principles is the Gestalt Laws of Perceptual Organization (see Wertheimer, 1934; Chapter 3, this volume). These laws dictate that distinct static perceptual elements will be seen as belonging to a single configuration if they are near one another ("proximity"), similar in terms of one or more visual dimensions ("similarity"), smooth continuations of one another ("good continuation") or parallel ("common fate") in the 2D plane. In terms of the visual description, these principles will determine how variables are linked via the "part" relation in structures like those in Figures 6.8a (where the law of similarity links asterisks to asterisks and circles to circles), 6.8b (where common fate links the asterisks to the line, and similarity links the asterisks to one another), and 6.8c (where good continuation keeps the straight and curved lines distinct, proximity links the asterisks and crosses to their respective lines, and similarity links asterisks to asterisks and crosses to crosses). Figure 6.8d shows how 6.8c would be represented in a visual description.

---

[3] This proposal is similar to Bertin's conjective that a focal percept (his "image", the content of a "perceptual glance") may consist of a spatial location plus the value of one "retinal dimension" at that location. It is not clear, however, why one should suppose that only one nonspatial dimension can be encoded at a given location. Indeed, in the discussion of coordinate systems for nonspatial dimensions, we discuss evidence that in fact several physical dimensions may be encoded simultaneously by the human visual system.

INSERT FIGURE 6.8 HERE

There is another way of indicating the effects of grouping within visual descriptions. That is to link each member of a group to every other member using either the relation that gave rise to the grouping, or simply the relation "associated with". Thus, the visual array in Figure 6.8a, above, could also be represented as in Figure 6.9:

INSERT FIGURE 6.9 HERE

This notation can be used to indicate that the variables are grouped together perceptually, but not so strongly as to be a distinct perceptual unit. In the rest of this chapter, we will use both notations for grouping, though no theoretical distinction will be implied by the choice.

C. Representation of Magnitude

Implicit in our earlier discussion of the psychophysics of visual dimensions was the assumption that these dimensions are represented by continuous interval scales in visual descriptions. Though the fine discriminations and smooth magnitude estimation functions found in psychophysical experiments strongly warrant this assumption, we have reason to believe that quantity can be mentally represented in other ways as well. First, there is evidence from experiments on the absolute identification of values on perceptual continua that people cannot remember verbal labels for more than about seven distinct levels of a perceptual continuum (Miller, 1956), and that in making rapid comparisons between remembered objects, subjects' reaction times are insensitive to the precise values of objects belonging to distinct, well-learned categories (Kosslyn, Murphy, Bemesderfer, and Feinstein, 1977). Findings like these suggests that quantity can also be represented (indeed, in memory must be represented, in certain circumstances) by one of a set of seven or so discrete symbols each specifying a portion of the range of quantities. These symbols could be signified by the Roman numerals I through VII.

Second, it is useful to distinguish between ratio values, where quantity is represented continuously but the units are arbitrary, and absolute values, where the units are well-defined. The perception of pitch is a notorious example where a precise mental representation of a dimension is possible, but where for a majority of people, no absolute units can be assigned to the stimuli. Length, on the other hand, is an example of a continuum which people can judge either in ratio terms (e.g., one object being 1.7 times as long as another), or in terms of the well-known inches-feet-yards scale (e.g., Gibson and Purdy, 1956). Indeed, whether subjects in magnitude estimation experiments are asked to use a well-learned versus their own arbitrarily-selected modulus for estimated magnitude apparently affects the resulting judgements

(Stevens, 1957). Thus, internal descriptions must discriminate between these two forms of magnitude, which we will refer to as "interval-value" and "absolute-value", though ordinarily, visual descriptions will only contain "interval-value" propositions.

Finally, as every commercial sign-maker can attest, values on a continuum that are extreme in comparison to values of that continuum for other objects in a scene are very likely to be perceptually encoded (as opposed to less extreme values, which are apt to be encoded only if attended to). To account for this salience principle, relatively extreme values will be represented redundantly in visual descriptions: in ordinary propositions such as "height (x, α)", as before, and also by special one-place predicates indicating the extremeness of the value along the particular dimension, such as "tall (x)", "bright (x)", "short (x)", etc. When capacity limitations of visual descriptions are discussed later in the paper, it will be assumed that these special predicates have a very high probability of being encoded in the visual description.

D. Coordinate Systems

To express a unidimensional quality like lightness, one need specify in advance only the origin and the units of the scale to be used. However, for objects that vary along a number of continua, like the position of an object on a two-dimensional piece of paper, or rectangles in a set varying in height and width, one has to specify how the variation will be partitioned into dimensions and how each dimension will be represented. This is the isue of which coordinate system is appropriate to represent an object in a set varying along several dimensions. In the case of dimensions that refer to spatial location, Bertin invents the term "construction schema" to refer to the way that the spatial dimensions of a graph are partitioned. This involves questions about whether a polar or a rectangular coordinate system is used, whether there is a single or multiple origins, and so on. In the case of nonspatial dimensions like color or shape, Bertin does not use the vocabulary of coordinate systems, but it is equally appropriate. We will briefly discuss some of the considerations relevant to the choice of coordinate systems for multidimensional stimulus, separately for nonspatial and spatial (more exactly, nonpositional and positional) dimensions.

1) Nonspatial dimensions. Many visual objects can logically be parameterized in more than one way. For example, rectangles can be classified by their heights and widths, or by their sizes and shapes (where "shape" could be a dimension ranging from "very tall and narrow" through "square" to "very short and wide"). Similarly, colors can be represented by their hues, saturations, and brightness (e.g., blood is saturated with a dominant wavelength of 700 nm, and boiled shrimp are desaturated with a dominant wavelength of 700 nm), or by their closeness to various "focal colors" (Rosch, 1975; here blood might be highly crimson and not very pink, whereas boiled shrimp would be highly pink and not very crimson).

One might expect there to be perceptual consequences of which set of dimensions a stimulus is encoded along, and indeed there are. Garner (1974, see also Chapter 3) distinguishes between separable and integral perceptual dimensions. According to Garner, each of a pair of "separable" dimensions may be attended to independently of the other, whereas one cannot attend to one member of a pair of integral dimensions without attending to the other as well (see Chapter 3 for a discussion of the experimental procedures used to ascertain whether a given pair of dimensions is separable or integral). One way of translating Garner's terminology into our own is to consider separable dimensions to be those physically defined dimensions that are also psychological dimensions. That is, if color and size are found to be separable dimensions in attention tasks, we may infer that humans in fact encode objects into their color and into their size, recording both dimension values as separate parts of their mental representations of the objects. Selective attention is possible because the dimensions *are* separately represented internally; one can be processed while the other remains in storage. Thus, the separability of two physical dimensions is prima facie evidence that those dimensions are the ones used in the mental representation.

Integral dimensions, on the other hand, may very well be psuedo-dimensions, psychologically speaking: the reason that humans apparently cannot ignore the height of a rectangle while attenidng to its width is that height and width are not the dimensions that humans, left to their own devices, would encode into their mental representations of a rectangel, Rather, the psychologically-relevant dimensions might be *size* and *shape*, in terms of, say, a *fatness - skinniness* dimension. When asked to attend to the height of a rectangle, there would be no parameter or symbol in the mental encoding of the rectangle that represents height alone and thus that can be processed while other parameters are left alone. Rather, both the size and the fatness-skinniness parameter implicitly contain information about height, and both would have to be processed so that their values may be transformed into the height value that performance on task demands. This transformation process could account for the increase in time required to sort stimuli along integral versus separable dimensions. Similarly, the reason that humans apparently cannot ignore the saturaiton of a color while attending to its hue may be that the color is not naturally encoded into separate hue and saturation parameters, but into parameters representing its proximity to various focal colors such as pink, red, brown, and so on.

In sum, we may determine exactly which dimensions humans use in their mental representations of multidimensional stimuli by examining the results of

Garner-type experiments. If a pair of physically-specified dimensions is separable, we may conclude that there is a mental parameter corresponding to each of those dimensions. On the other hand, if a pair of dimensions turns out to be integral, we may conclude that the mental parameters representing those stimuli correspond to a different dimensionalization of the stimuli from the one the experimenter had in mind. Intermediate cases (e.g., where no possible dimensionalization of a stimulus set yields perfect separability) may reflect multiple parameterized encodings of a stimulus, the various encodings differing in strength of activation (see the section on Processing Constraints below).

2) Spatial Dimensions. In their influential paper on shape recognition, Marr and Nishihara (1977) proposed that memory representations of shape are specified with respect to object-centered cylindrical coordinate systems. Furthermore, the coordinate systems are distributed: instead of there being a global coordinate system with a single origin and set of axes, there is a cylindrical coordinate system centered on the principle axis of the object (e.g., in the case of an animal, its torso), in which is specified the origins and axes of secondary coordinate systems centered on the various parts of the object attached to the principle axis (e.g., the animal's head and limbs). These secondary coordinate systems are, in turn, used to specify the origins and axes of smaller coordinate systems centered on the constituent or attached parts of the secondary part (e.g., the thigh, shin, and foot of the leg), and so on. We will adopt here the following aspects of Marr and Nishihara's theory: 1) shapes and positions are mentally represented principally in polar or rectangular coordinates (the former is just a slice of a cylindrical coordinate system orthogonal to its axis; the latter is just a slice of a cylindrical coordinate system including its axis). 2) The locations of the different elements of a scene are represented in separate, local coordinate systems centered upon other parts of the scene, not in a single, global coordinate system. This means that in the visual description, the specification of

locations (and also of directions and of parameterized shapes) of objects will be broken down into two propositions, one specifying the object upon which the coordinate system will be centered, the other specifying the extent or value of the object within the coordinate system, as in Figure 6.10.

INSERT FIGURE 6.10 HERE

In fact, it is generally more perspicuous to indicate the extent along each dimension, and the location of the axis of the coordinate system corresponding to that dimension, separately, as in Figure 6.11

INSERT FIGURE 6.11 HERE

The important question of which objects may serve as the coordinate system for which other objects has received little attention in the vision literature, but the following condition seems to be a plausible first approximation: the location (or direction, or shape parameters) of object _a_ will be mentally specified in a coordinate system contered on object _b_ when: 1) _b_ is larger than _a_, and 2) _a_ and _b_ are perceptually grouped according to one or more of the Gestalt laws.

Processing Constraints on Visual Descriptions

Since, with deliberate effort, people can probably encode an unlimited number of properties (e.g., the angle formed by imaginary lines connecting a standing person's right thumbnail, navel, and right kneecap), visual descriptions can in principle be arbitrarily large. In practice, however, two factors will limit the size of visual descriptions:

1) Processing Capacity. Most models of cognitive processing have restrictions on the capacity to maintain the activation of nodes in a short-term visual description (Anderson & Bower, 1973; Newell & Simon, 1973). Specifically, it is claimed that between 4 and 9 nodes may be kept active at one time (see Chapter 2), fewer if processing resources are being devoted to some concurrent task. This limitation reflects the well-known finiteness on human immediate memory and processing capacity.

2) Default Encoding Liklihood and Automaticity. As mentioned, any predicate in a person's visual repertoire can be added to a visual description in response to higher-level processes testing for the presence of a particular predicate applied to a particular variable (e.g., "is x a square?"). However, before these top-down processes come into play, a number of predicates will be assembled into a visual description, because they are "just noticed". Different predicates have different probabilities of being encoded under these "default" circumstances. Presumably, some predicates innately have a high default encoding likelihood [e.g., enormous (x), dazzling (x)] whereas the default encoding likelihood of others is determined by familiarity and learned importance. Shiffrin and Schneider (1971) and Schneider and Shiffrin (1977) propose that when a person frequently assigns a visual pattern into a single category, he or she will come to make that classification "automatically", that is, without the conscious application of attentional or processing capacity. Translated into our vocabulary, this means that frequently-encoded predicates will have a high default encoding likelihood. A number of experiments applying Shiffrin and Schneider's proposals to the learning of visual patterns confirm that the recognition of patterns becomes rapid, error-free, and relatively insensitive to other attentional demands as the patterns become increasingly well-practised.

Therefore, it is important to distinguish among several sizes of visual descriptions. A description that is assembled automatically by purely data-driven (as opposed to top-down or conceptually-driven) encoding processes will be called the "default visual description". Its composition will be determined by the relative "default encoding likelihoods" of the various predicates satisfied by the visual array. In contrast, a description that is shaped by conceptual processes testing for the presence of visual predicates at particular locations in the array will be called an "elaborated visual description". Visual descriptions can also be classified in terms of whether short-term memory limitations are assumed to be in effect. A small visual description such as can be activated at a given instant will be called the "reduced visual description"; a visual description that includes all the predicates whose default encoding likelihoods are above a certain minimum, plus, all the predicates that are successfully tested for by top-down processes, will be called the "complete visual description". The complete visual description will correspond to the description encoded by a hypothetical graph reader with unlimited short-term memory, or to the description integrating the successive reduced descriptions encoded by a normal graph reader over a long viewing period. One way to think quantitatively of the size of the default visual description that a person will encode is to suppose that the probability of a given true predicate's entering into a visual description is a function of its default encoding likelihood multiplied by a constant between zero and one corresponding to the amount of capacity available (i.e., not devoted to other concurrent tasks). When the constant is one, the resulting description will be a "complete" visual description; as the constant decreases with decreasing available processing capacity, the size of the description will be reduced accordingly. We adopt the final assumption that the level of activation of a node begins to decrease steadily as soon as it is activated, but that the reader can repeatedly re-encode the description by reattending to the graph (see the voluminous literature on decay and rehearsal in short-term memory summarized, for example, in Crowder, 1975). Since encoding is probabilistic, the description will differ in composition somewhat from one encoding to the next.

V. An Example

Now that we have some constraints on the size and composition of visual descriptions, we can examine how a particular graph might be described mentally. This will be the first step in working through an example of how a graph is understood according to the current theory. The example, shown in Figure

6.12, is a bar graph plotting the price per ounce of a precious metal we will call "graphium" over a six month period.

INSERT FIGURE 6.12 HERE

A "complete" default visual description is shown in Figure 6.13. (Dotted lines represent propositions, omitted for the sake of clarity, that may be deduced from nearby propositons for similar parts.)

INSERT FIGURE 6.13 HERE

Most aspects of this visual description are motivated by the constraints outlined in the previous secton. The scene is parsed into subscenes, each occupying a distinct location in the visual array (though for readability's sake, the locations for the subscene nodes will not always be printed in the future). This parse is done according to the Gestalt principles, yielding separate nodes for the "L"-shaped framework and for the group of bars. By those same principles, the framework is connected by the "part" predicate to nodes representing its vertical and horizontal segments, and each of these is linked by "near" predicates to nodes representing the conceptual meaning of that text. Of course, the meaning of expressions like "price of graphium" is, in all likelihood, mentally represented by an assembly of nodes linked in complex ways to the nodes representing the visual appearance of the text, but since the process of reading text is not our concern here, this simplified notation will suffice (the predicate associated with these "meaning" nodes will be replaced within quotation marks to indicate that they are not in fact unitary predicates). Predicates for the "bar" shape are attached to each bar node; the "tall" predicate is attached to the salient tallest bar; a pair of particularly discrepant bars is connected by the predicate "taller-than"; and the set of four progressively shorter bars is grouped together under its own node with its own shape predicate "descending-staircase." Finally, the height and horizontal position of each bar is specified with respect to a coordinate system centered on the appropriate framework segment, due to the framework's being larger than the bars and associated with them by proximity and common fate.

VI. Conceptual Messages, Conceptual Questions

We now have an example of the immediate input to the graph comprehension process. Before specifying that process, it would be helpful to know what its output is as well. One can get a good idea of what that output must be simply by looking at a graph and observing what one remembers from it in the first few moments of seeing it or after it has just been removed from view. In the case of the graph in Figure 6.12, one might notice things like the following: a) the price of graphium was very high in March; b) the price was higher in March than in the preceeding month; c) the price steadily declined from March to June; d) the price was $20/ounce in January; e) the price in June was x (where x is a mental quantity about half of that for January, about a fifth of that for May, etc.). Basically, we have a set of paired observations here, where the first member can be a particular value of the independent variable (e.g., "March"), a pair of values (e.g., "March vs. February"), or a range of values (e.g., "the last four months"). The second number of each pair can be a ratio value (e.g., a value *x* along some mental ratio scale), an absolute value (e.g., "$20/ounce"), a difference (e.g., "larger"), a trend (e.g., "decreasing"), or a level (e.g., "high"). Bertin first pointed out these options, using the term "elementary questions" for those referring to single values, "intermediate questions" for differences, and "superior questions" for trends. This information can be expressed in a representation consisting of a list of numbered entries, each specifying a pair (or, for more complex graphs, an n-tuple) of variables, the extent or type of each independent variable (e.g., ratio-value, pair, range), and the value (or difference or trend) of the corresponding dependent variable. Thus, the conceptual message representing the information which we are assuming has been extracted from the graph in Figure 6.12 will look like this (the intuitive meaning of each entry can be made clearer by assuming the entry is a sentence beginning with the word when):

1: V_1 absolute-value = March, V_2 level = high
2: V_1 pair = March & February, V_2 difference = higher
3: V_1 range = March - June, V_2 trend = decreasing
4: V_1 absolute-value = January, V_2 absolute-value = \$20/oz.
5: V_1 absolute-value = June, V_2 ratio-value = x.

In general, conceptual messages will be of the following form:

i: V_a ratio-value $= \alpha$, V_b ratio-value $= \beta, \ldots$
or or
absolute-value absolute-value
or or
pair pair
or or
range range

i designates the ith of an arbitrary number of entries (in principle), V_a designates the ath of an arbitrary number of variables, and α designates a specific value in a form appropriate to the entry (e.g., a "higher" or "lower" primitive symbol if the entry specifies a difference between values of the second variable corresponding to a pair of values of the first)[4]. Note that the variables are differentiated by subscripts instead of being named by their real-world referents (e.g., "month"); this was done in recognition of people's ability to extract a great deal of quantitative and qualitative information (indeed, virtually the same information) when a graph has no labels at all, leaving the referents of the variables unknown. When the referents are known, the conceptual message can indicate this with entries like the following:

6: V_1 = months, V_2 = price-of-graphium.

---

[4] It is possible to have several equations in an entry refer to the same variable, eg.:

17: V_1 absolute-value = 14, V_1 ratio-value = 132, V_1 level = high, V_2 level = low.

Presumably, when the reader has integrated all the information he or she wishes to extract from the graph, he or she can make the message representation more economical by replacing each $\underline{V}_i$ by its associated referent symbol.

From here, it is a simple matter to devise a notation for conceptual questions. (Recall that a conceptual question is a piece of information that the reader desires to extract from a graph.) One can simply replace the $\underline{\alpha}$ or $\underline{\beta}$ in the generalized entry presented above by the "?" symbol, indicating that that is the unknown but desired information. Thus, if a person wishes to learn the price of graphium during the month of April, we posit that he or she has activated the representation

7: V_1 absolute-value = April, V_2 absolute value = ?.

If the reader wishes to learn the trend of graphium prices during the first two months, he or she sets up the representation

8: V_1 range = January-February, V_2 trend = ?.

If the reader wishes to learn the month in which graphium prices were low, he or she activates

9: V_1 absolute-value = ?, V_2 level = low,

and so on.

VII. The Graph Schema

So far, our theory has implicated an information flow diagram like the one in Figure 6.14.

INSERT FIGURE 6.14 HERE

Now, we must specify the unknown component labelled with a "?". From the flow chart, we can see what this component must do: 1) it must specify how to translate the information found in the visual description into the conceptual message, and 2) it must specify how to translate the request found in a con ceptual question into a process that accesses the relevant parts of the visual description (culminating as before in one or more entries in the conceptual message). Furthermore, since (1) and (2) will involve different sorts of translations for different types of graphs (e.g., for line graphs versus bar graphs), the unknown component will also have to 3) recognize which type of graph is currently being viewed. The structure that accomplishes these three tasks will be called a graph schema, and it, together with the processes that work over it, will be discussed in this section.

A. Schemas

A schema is a memory representation, embodying knowledge in some domain, consisting of a description containing "slots" or parameters for as yet unknown information. Thus, a schema can specify both the information that must be true of some represented object of a given class, and the sorts of information that will vary from one exemplar of the class to another (For detailed presentations of various schema theories, see Minsky, 1975; Winston, 1975; Norman & Rumelhart, 1975; Bregman, 1977; Schank & Ableson, 1977). To take a simple example unrelated to graphs, Figure 6.15 could be a schema for telephone numbers, specifying the number and grouping of the digits for any number but not the identity of the digits for any particular number, these being represented by the parameters A-J.[5]

INSERT FIGURE 6.15 HERE

This schema can be instantiated for a given person, becoming a representation of his or her particular telephone number, by replacing the parameters labeling the lowermost nodes by actual numerical predicates. In doing so, one is using the schema to recognize a candidate character string as a telephone number, by matching the schema against a visual description of the candidate string. The visual description of an as yet unrecognized number will be identical to the schema, except that it lacks the conceptual nodes like "area code" and "exchange" and that it contains constants in place of parameters. Once the schema is instantiated by the visual description, one can use it to retrieve desired information about the telephone number using a node-by-node net searching procedure (i.e., one can quickly find "the first digit of the exchange" without searching the entire string, by starting at the top node and following the appropriate arrows down until the bottom node labeled by the desired number is reached). The double labeling of nodes is what allows schemas to be used both for recognition and for searching: a visual description of a to-be-recognized pattern will contain labels like "digit", but not "area code", so the "digit" labels in the schema are necessary for recognizing the object. However, the search procedures will be accessing conceptual labels like "area code", so these are necessary, too.

---

[5] These upper case parameters, which stand for unknown predicates, should not be confused with lower case variables, which stand for perceptual entities and correspond to nodes in the visual description (although usually, the variable itself is omitted and only the node is depicted).

B. Graph Schemas: A Fragment

It seems, then, that a schema of this sort for graphs might fulfill two of our three requirements for graph knowledge structures: recognizing specific types of graphs, and directing the search for desired pieces of information in a graph. What we now need is some device to translate visual information into the quantitative information of the type found in the conceptual message. These devices, which we will call message flags, consist of conceptual message equations, usually containing a schema parameter, which are appended to predicates (nodes or arrows) in the graph schema. When such a node or arrow is instantiated by a particular visual description for a graph, the parameters in the message flag are replaced by the corresponding value in the instantiated schema, and the equation is added to the conceptual message. Figure 6.16(a) illustrates equation flags for a fragment of a bar graph schema (the flags are enclosed in rectangles, and are attached to the nodes they flag by dotted lines).

INSERT FIGURE 6.16 HERE

When a reader encounters the graph represented by the fragment of a visual description in Figure 6.16b (the numbers representing values along a mental ratio scale with arbitrary units), he or she can instantiate the schema (i.e., replace the parameters A and B by the values 4 and 37), and add an entry to the conceptual message. All equations sharing a given i prefix are merged into a single entry, and each i is replaced by a unique integer when the entry is added to the conceptual message. Thus, the following entry is created:

1: V_1 ratio-value = 4, V_2 ratio-value = 37

This informal sketch should give the reader a general idea of how the graph schema is used in conjunction with the visual description to produce a conceptual message. In the sections following, we present a comprehensive bar graph schema, and define more explicitly the processes that use it.

C. A Bar Graph Schema

Figure 6.17 presents a substantial chunk of a schema for interpreting bar graphs. It is, intentionally, quite similar to the visual description for a bar graph in Figure 6.13. The graph is divided into its L-shaped framework and its specifier material, in this case, the bars. The framework is divided into the abscissa and the ordinate, and each of these is subdivided into the actual line and the text printed alongside it. In addition, the "pips" cross-hatching the ordinate, together with the numbers associated with them, are listed explicitly. The height and horizontal position of each bar are specified with respect to coordinate systems centered on the respective axes of the framework, and each bar is linked to a node representing its nearby text. An asterisk followed by a letter inside a node indicates that the node, together with its connections to other nodes, can be duplicated any number of times in the visual description. The letter itself indicates that each duplication of the node is to be assigned a distinct number, which will appear within the message flags attached to that instance of the node.

INSERT FIGURE 6.17 HERE

The message flags specify the conceptual information that is to be "read off" the instantiated graph schema. They specify that each bar will contribute an entry to the conceptual message. Each entry will equate the ratio value of the first variable (referred to in the description as "IV", for Independent Variable) with the horizontal position of the bar with respect to the abscissa, and will equate the second variable (the "DV", or Dependent Variable) with the bar's height with respect to the ordinate. In addition, the absolute value of the independent variable for an entry will be equated with the meaning of whatever label is printed below it along the abscissa. Finally, the referents of each variable will be equated with the meaning of the text printed alongside its respective axis.

In devising these formalisms, we were at one point distressed that there was no straightforward way to derive absolute values for the dependent variable. The ratio value of each bar, corresponding to its height, could easily be specified, but since the absolute values are specified in equal increments along the ordinate, far from most of the bars, and specific to none of them, no simple substitution process will do. However, a simple glance at a bar graph should convince the reader, as it convinced us, that this is not a liability but an asset. The absolute value of the dependent variable at a given level of the independent variable is indeed not immediately available from a bar graph.

Instead, one seems to assess the height of a bar in terms of some arbitrary perceptual or cognitive scale, and then search for the pip along the ordinate whose vertical position is closest to that height. The number printed next to that pip, or a number interpolated between the numbers printed next to the two nearest pips is deduced to be its absolute value. In contrast, the absolute value of a given level of the independent variable (i.e., which month it is), or the relative values of the dependent variable (e.g., its maximum and minimum values, its trends, or differences between adjacent values) seem available with far less mental effort. The most natural mechanism for representing absolute values of the dependent variable within the bar graph schema, and the one that happens to be in accord with the actual difficulty of perceiving these values, is to add to the conceptual message special entries asserting an equivalence between a certain level of the referent's absolute value and a certain level of the referent's ratio value, each entry derived from a labeled pip on the ordinate. The leftmost message flag in Figure 6.17 sets up these entries; the symbol "=" indicates that the two equations are equivalent. Presumably, higher-level inferential processes, unspecified here, can use these equivalence entries to convert ratio values to absolute values within other entries in the conceptual message, calculating interpolated values when necessary.[6]

Earlier, we mentioned that the visual system can encode predicates that stand for well-defined groups of objects, and also that conceptual messages can contain entries specifying a trend of one variable over the range of another. An implication of the theory, then, is that graph readers (or at least experienced graph readers) should be able to translate directly a higher-order perceptual pattern, such as a group of bars comprising a staircase, into the quantitative trend that it symbolizes, without having to compute the trend by successively examining each element. Furthermore, the difference in height between a pair of adjacent bars might be encodable into a single predicate, which should be directly translatable into an entry expressing a difference in the symbolized values. Also, a salient perceptual entity might be encoded as extreme (independently of the encoding of its precise extent on a ratio scale), and this should be directly translatable into an entry expressing the extremeness of its corresponding variable value, again without the mediation of ratio scale values. These direct translations, which, as we shall see, play an important role in predicting the difficulty of a graph or the effectiveness of a graph reader, are accomplished by the message flags in Figure 6.18 (which should actually be part of Figure 6.17, but is depicted separately for the sake of clarity). Figure 6.18 shows that bars in a graph can be described not only in terms of their heights and horizontal positions, but also in terms of being extremely tall or short, in terms of differences between the heights of adjacent pairs, or in terms of groups that constitute a perceptual whole. In each case the appropriate equation is attached to the predicate which encodes the attribute. Two additional notational conventions are introduced in the figure: the location of a pattern that occupies an extended region of the array is specified by its endpoints along a ratio scale (i.e., "H-I"), both in the visual description and in the conceptual message. In addition, one of the equation flags for a pair of bars makes reference to nodes standing for the bars themselves, p_j and q_j, rather than for an attribute like horizontal position. It is assumed that when a pair of bars is encoded as a pair, some information

---

[6] The schema presented here perhaps unfairly anticipates that the bar graph example will have individual labels for each bar along the abscissa, and a graduated scale along the ordinate. In fact, graduated scales often appear along the abscissas of bar graphs as well. In a more realistic bar graph schema, the subschema for the pips of a graduated scale would be appended to the abscissa as well as to the ordinate.

INSERT FIGURE 6.18 HERE

about each bar is encoded as well. This information, be it ratio value, absolute value, or level, can then be linked with or substituted for appropriate symbols for the bars (p_j or q_j) within the entry for the pair.

VIII. Processes

In the account so far, we have relied upon the intelligence and cooperativeness of the reader to deduce how the various structures are manipulated and read during graph comprehension. In order to use the theory to make predictions, it will be necessary to define explicitly the procedures that access the structures representing graphic information. Four procedures will be defined: a MATCH process that recognizes individual graphs as belonging to a particular type, a message assembly process that creates a conceptual message out of the instantiated graph schema, an interrogation process that retrieves or encodes new information on the basis of conceptual questions, and a set of inferential processes that apply mathematical and logical inference rules to the entries of the conceptual message.

A. The MATCH Process

The term is borrowed from Anderson and Bower's (1973) theory of long-term memory. This process compares a visual description in parallel with every memory schema for a visual scene, computes a goodness-of-fit measure for each schema (perhaps the ratio or difference between the number of matching nodes and predicates and the number of mismatching nodes and predicates), and selects the schema with the highest goodness-of-fit measure. This schema, or rather, the subset of the schema that the limited capacity processes can keep activated, is then instantiated (i.e. the parameters in the schema are replaced by the appropriate constants found in the visual description.). This is the procedure alluded to in vague terms before, that uses the graph schema to recognize a graph as being of a certain type (e.g., bar graph, pie graph).[7]

B. Message Assembly

This process accomplishes the translation from visual information to conceptual information, also alluded to in previous sections. It searches over the instantiated graph schema, and when it encounters a message flag, it adds the message it contains to the conceptual message, combining into a single entry all equations sharing a given prefix (i.e., all those beginning with the same "i:"). It is assumed that at the time that the MATCH process instantiated the parameters of the graph schema, the parameters within the message flags were instantiated as well.

Memory and processing limitations imply that not every message flag in the graph schema is converted into an entry in the conceptual message: some may not be instantiated because the visual description was reduced, or because the default encoding likelihood of the predicate was low; some may not be instantiated because of noise in the MATCH process; and some may be skipped over or lost because of noise in the message assembly process. For these reasons, we need a process that adds information to the conceptual message in response to higher-level demands.

---

[7] This process has been oversimplified in several ways, in accordance with certain oversimplifications in the graph schema itself. For one thing, conceptual labels like "abscissa" do not appear in visual descriptions, and so should not count in the goodness of fit calculations. This could be accomplished by distinguishing the conceptual or graph-specific predicates from the rest, perhaps by listing them, too, as message flags, which are "read off" the schema, but not used to instantiate it. The second complication is that different nodes and predicates should count differently in the recognition process. Some might be mandatory, some might be mandatorily absent, some might be characteristic to various degrees, some might occur in sets from which one member must occur, and so on. There are several ways of accomplishing this, such as the introduction of logical operators into schemas, or the use of a Bayesian recognition procedure, but limited space prevents us from outlining them here (see Anderson, 1976; Anderson & Bower, 1973; Minsky, 1975; Winston, 1975; Smith, Shoben, & Rips, 1973).

C. Interrogation

This process is called into play when the reader needs some piece of information that is not currently in the conceptual message (e.g., the difference between two values of the dependent variable corresponding to a given pair of independent variable values). As mentioned, each such request can be expressed as a conceptual message entry with a "?" replacing one of the equation values. The interrogation process works as follows: the message flag within the graph schema that matches the conceptual question (i.e., is identical to it except for a constant or parameter in the place of the "?") is activated. If it already contains a constant (i.e., if the equation it contains is instantiated, and thus, complete), the equation is simply added to the conceptual message. If it contains a parameter (i.e., is incomplete), the part of the visual description that corresponds to that branch of the schema is checked to see if it contains the desired constant (e.g., if a certain ratio-value of the dependent variable is desired, the visual description is checked for the presence of a constant attached to the node representing the bar's height). If this constant is absent from the visual description, the encoding process for the relevant predicate (e.g., the process that encodes height) is commanded to retrieve the desired information for the relevant part in the visual array. It can do so by using the retinal coordinates attached to the node for the part which are assumed to be present in the visual description (though they have been omitted from the diagrams in this chapter). Often, however, these coordinates will have decayed, and the coordinates of an associated part together with the degree and direction of the association will be used to direct the encoding process to the correct location in the visual array. In other words, the conceptual question can initiate a top-down search for the desired part or part parameter in the array. Once the desired information is encoded into the visual description, it can be instantiated in the schema and its message flags, and the instantiated equation within the flag can be added to the conceptual message.

D. Inferential Processes

Human intelligence consists of more than the ability to read graphs. In the category *inferential processes*, we include the ability to perform arithmetic operations on the quantitative information listed in the conceptual message (e.g., calculating the rate of increase of a variable by subtracting one value from another value and dividing by a third value), to infer from the context of the graph (e.g., the paragraph in which it is embedded) what information should be extracted from the graph, to draw qualitative conclusions relevant to some domain of knowledge based on the information in the graph, and so on. Naturally, we have little to say about these abilities; they are part of the study of cognition in general and not the study of graph comprehension. However, we mention them here because many types of information can be obtained either directly from a conceptual message or indirectly from inferential processes operating on the conceptual message. Which method is used, we shall see, affects the difficulty of a graph and the efficiency of a graph reader.

The flow of information specified by the current theory is summarized in Figure 6.19, where blocks represent information structures, and arrows represent processes that transfer information among them.

INSERT FIGURE 6.19 HERE

IX. Where do Graph Schemas Come From?

The graph schema discussed so far embodies knowledge of bar graphs (in fact, a subset of bar graphs). Clearly, the theory must also account for people's ability to read other common types of graphs (line graphs, pie graphs, pictograms, etc.) and to understand completely novel forms of graphs as well (e.g., one in which the length of a ray of light emitted from a disc represents the price of gold in a given month). We propose people create schemas for specific types of graphs using a *general graph schema*, embodying their knowledge of what graphs are for and how they are interpreted in general. A plausible general graph schema is shown in Figure 6.20. There are three key pieces of information contained in the schema. First, some objects, or parts of objects (specifier material) are described in terms of several visual attributes. Each visual attribute symbolizes a conceptual variable, and the set of values of the *n* visual attributes encoded for an object or object part corresponds to a particular n-tuple of associated values of the respective conceptual variables for a given conceptual entity. Second, the ratio magnitudes of attributes are usually to be specified in terms of a coordinate system centered upon a part of the graph framework. Third, textual material perceptually grouped with an object specifies the absolute value of the object; textual material perceptually grouped with the framework specifies the real-world referent of the attribute that the coordinate system centered on the framework helps to specify; textual material associated with specific local regions of the framework specifies pairings of absolute and ratio values of the attribute specified by the associated coordinate system. In other words, the general graph schema encodes knowledge of graphs in a way that

INSERT FIGURE 6.20 HERE

respects the basic assumptions underlying our analytic scheme (Chapter 2), in which graphs are parsed into specifier material, a framework, and a set of labels. Note that for maximum generality, text is linked to perceptual entities by the predicate "associated", which can symbolize proximity, similarity, continuity, and so on. This helps to encompass graphs with parts directly labelled, and graphs exploiting common colors or shapes in keys and legends. Similarly, the predicate "attribute" is meant to encompass length, width, orientation, lightness, color, etc. However, the indispensibility of visual space motivates "geometric shapes" as opposed to arbitrary visual predicates being specified as typical frameworks, and spatially localizable "parts" being specified as the units over which attributes are defined.

In encountering a certain type of graph for the first time, a reader will generate a specific graph schema for it using the general graph schema. The reader will have to replace the predicates "specifier material", "associated", "attribute", "geometric figure", and so on, by the actual visual predicates found in the visual description of the novel graph. This will be possible when the visual description has a structure similar to that of the general graph schema, with objects described in terms of attributes defined with respect to a framework, and textual labels associated with each. In addition, an astute graph reader will add to the new specific graph schema higher-order predicates (e.g., "descending-staircase") that can be taken to symbolize global trends (e.g., a decrease in the dependent variable). However, the availability of these higher-order predicates, and how transparently they symbolize their trends, will differ arbitrarily from graph type to graph type, and so these predicates cannot be included in any simple way within the general graph schema but must be created case-by-case. This process will be discussed in more detail in the section describing what makes a graph reader efficient.

Pushing the question back a step, we may ask, "Where does the general graph schema come from?" This question is more profound, and the answer to it is correspondingly murkier. In one sense, one could answer that people are explicitly taught how to read certain types of graphs. But, this still leads one to wonder how people can generalize from the small set of graph types that they are exposed to in school (basically, bar graphs, line graphs, pie graphs, and pictograms) to the myriad exotic forms that are created and easily understood in, say, TIME magazine. This is especially problematic given that formal instruction in graph reading does not teach the abstract concepts such as "attribute", "extent", "ratio value", and so forth, that in fact define what all graphs have in common. A deeper answer to this question, then, would seem to lie in a basic human ability to associate a scale of values (i.e., an attribute with an "extent" predicate) in one domain with a scale of values in virtually any other domain, so long as the "positive" end of one scale, as mentally represented, coincides with the "positive" end of the other. Thus, there are lawful relations governing such diverse phenomena as the order of words with different sounds in conjoined phrases, the choice of which member of a pair of associated symbols or metaphors will represent specific ideas, and which way of installing a switch or gauge will yield the most efficient man-machine interaction. See Cooper and Ross (1975), and Pinker and Birdsong (1979) for discussions of some of these principles and their significance.

X. The Difficulty of Comprehending a Graph

In this section, we consider what makes different types of graphs easy or difficult when particular types of information have to be extracted (by "type of information", we are referring to different conceptual questions, such as ones referring to ratio values vs. differences vs. trends.).

Aside from the limitations of the peripheral encoding mechanisms (i.e., limits on detectability, discriminability, and the accuracy of encoding magnitudes), the structures and processes described here permit any quantitative information whatsoever to be extractible in principle from a graph. This is because no information is necessarily lost from the visual description "upward", and because there are no constraints on what the inferential processes can do with the information in the conceptual message.

In practice, though, limits on short-term memory and on processing resources will make different sorts of information easier or more difficult to extract. We have assumed that the visual description that is encoded is, in fact, is a small subset of the complete visual description, and that noise in the MATCH and message assembly processes causes only a subset of that reduced visual description to be translated into conceptual message information. The remaining conceptual message entries will contain the information that is "easily extracted" from a graph, since a simple lookup procedure suffices to retrieve the information. On the other hand, if the desired information is not already in the conceptual message, it will have to be generated either by the top-down interrogation process, which adds entries to the conceptual message, or by the inferential processes, which perform computations on existing entries. Each of these processes can involve a chain of (presumably) capacity-limited computations, and each process properly includes the lookup of information from the conceptual message. Therefore, they are necessarily more time-consuming and memory-consuming (since the results of intermediate computations must be temporarily stored) than the lookup of existing information in the conceptual message. And, in a limited-capacity, noisy system like the human mind, greater time and memory requirements imply increased chances of errors or breakdowns, hence, increased difficulty. We can call this conclusion the Graph Difficulty Principle: A particular type of information will be harder to extract from a given graph to the extent that inferential processes and top-down encoding processes, as opposed to conceptual message lookup, must be used.

This kind of principle is to be distinguished from the "operating principles" introduced earlier: those principles specified properties of the display itself which must be respected if a graph is to communicate effectively. This kind of principle explains why those particular stimulus properties have the effects they do; namely because of the structure and limitations of the human visual information processing system. In particular, the effects of violations of the earlier principles of discriminability, distortion, gestalt organization, integral/separable dimensionality and limited capacity (unit binding), all can be understood by reference to unfortunate effects on visual descriptions, and the effects of the remaining operating principles (with the exclusion of the two formal principles, which are not grounded in the properties of information processing per se) can be understood by reference to problems in matching description and stored schemata and/or deficiencies in the schema themselves.

There will, in turn, be two factors influencing whether a desired type of information (i.e., the answer to a given conceptual question) will be present in a conceptual message. First, a message entry will be assembled only if there are message flags specific to that entry appended to the graph schema. That, in turn, will depend on whether the visual system encodes a single visual predicate that corresponds to that quantitative information. For example, we have assumed that a bar graph schema appends message flags to predicates for height, horizontal position, extremeness in height, extreme differences in height between adjacent objects, and extended increases or decreases in height. This respectively makes ratio values of the dependent and independent variables, extremeness in value, extreme differences in values, and global trends easily extractible. On the other hand, there is no visual predicate for an object being a given number of ordinate scale units high, or for one bar's height to be a precise ratio of the height of another, or the leftmost and rightmost bars to be of the same height, and so on; therefore, there can be no message flags for and no conceptual message entries for the absolute value of the dependent variable, the exact ratio of dependent variable values corresponding to successive values of the independent variable, or the equality of dependent variable values corresponding to the most extreme independent variable values. If a reader wishes the graph to answer these conceptual questions, he or she can expect more difficulty than for the conceptual questions discussed previously.

The second factor influencing whether a conceptual message entry will be assembled is the encoding likelihoods of the predicates attached to the corresponding equation flags in the graph schema. In the example we have been using, if the predicate "descending-staircase" has a very low default encoding likelihood, and hence is absent from the visual description on most occasions, the entry specifying a decreasing trend will not find its way into the conceptual message until interrogated explicitly. Incidentally, apart from innateness and automaticity factors, the encoding likelihood of a predicate may also be influenced by "priming": when a graph schema is activated (i.e., when the graph is recognized as being of a particular type), the encoding likelihoods of the visual predicates are temporarily enhanced or "primed" (see Morton, 1969). In other words, when a graph is recognized on the basis of partial recognition, the schema makes the rest of the information more likely to be encoded for as long as the schema is activated.

As simple as the Graph Difficulty Principle is, it helps to explain a wide variety of phenomena concerning the appropriateness of different types of graphs for conveying different types of information. Consider Cartesian line graphs, for example. The English language has a variety of words to describe the shapes of lines: straight, curved, wiggly, V-shaped, bent, steep, flat, jagged, scalloped, convex, smooth, and many more. It also has words to describe pairs of lines: parallel, intersecting, converging, diverging, intertwined, touching, X-shaped, and so on. It is safe to assume that the diverse vocabulary reflects an equally or more diverse mental vocabulary of visual predicates for lines, especially since the indispensibility of visual space implies that predicates for configural spatial properties like shape should be readily available. The availability of these predicates affords the possibility of a line graph schema with a rich set of message flags for trends. For example, if "x" and "y" are nodes representing lines on a graph, with V_1 the abscissa, V_2 the ordinate, and V_3 the parameter, the propositions on the left side of Table 6.1 can be flagged with the conceptual message equations on the right side of the table:

Table 6.1

Predicate	Equation Flag
Flat (x)	V_2 trend = unchanging
Steep (x)	V_2 trend = increasing - rapidly
Inverted U-shape (x)	V_2 trend = quadratic
U-shape (x)	V_2 trend = quadratic
Jagged (x)	V_2 trend = random
Undulating (x)	V_2 trend = fluctuating
Straight (x)	V_2 trend = linear
S-shape (x)	V_2 trend = cubic
Rectilinaer (x)	V_2 trend = abruptly changing
Not flat (x)	V_1 affects V_2
Parallel (x,y)	V_1, V_3 additively affects V_2
Converging (x,y)	V_1, V_3 interactively affects V_2

This makes line graphs especially suited to representing particular trends of one variable over a range of a second, the covariation versus independence of two variables, and the additive versus interactive effects of two variables on a third. In contrast, the mental vocabulary for the shapes implicit in the tops of a set of grouped bars is poor, perhaps confined to "ascending-staircase", "descending-staircase", and "rectangular", as implied in Figure 6.18. Correspondingly, there will be fewer possibilities for specifying trends in a schema for bar graphs, and less likelihood of assembling specific "trend" and "affects" entries in the conceptual message when a bar graph is processed. And the predicates for a pair of shapes implicit in the respective tops of two integrated groups of bars will be even scarcer, preventing "additively affects" and "interactively affects" entries from being encoded. Small wonder, then, that line graphs are the preferred method of displaying multidimensional scientific data, where cause-and-effect relations, quantitative trends, and interactions among variables are at stake. To convince yourself of the appropriateness of line graphs for these purposes, try to determine the nature of the trend of V_2 over the range of V_1, and the nature of the interaction of V_1 and V_3 (a variable with two levels, A and B) on V_2, from Table 6.2, Figure 6.21a and Figure 6.21b.

Table 6.2

				V_2:		
	V_1:	1	2	3	4	5
V_3:	A	30.0	35.0	45.0	60.0	80.0
	B	20.0	32.0	45.0	57.5	70.0

INSERT FIGURE 21 HERE

It should be easy to see from the line graph in Figure 6.21b that at level A of Variable 3, Variable 2 is increasing and positively accelerating, whereas at B, it is increasing linearly. Similarly, one can see that Variables 1 and 3 interact in their effects on Variable 2. This is because the "straight" and "concave-up" predicates, corresponding to "linear" and "positively accelerating" trends, are readily encodable. In contrast, the like-colored bars in Figure 6.21a do not form a group where relative heights can be described by a single predicate, and so inferring the trend necessitates a top-down bar-by-bar height comparison, a difficult chore because it is hard to keep the heights of all the bars in mind (i.e., activated in the visual description) at once. It is even more difficult to extract the trends from the table, because not only is a number-by-number comparison necessary, but the process of encoding a multi-digit numeral's magnitude seems to be intuitively slower and more effortful than the encoding of a bar's height.[8,9]

However, try to answer the following question by examining the table, bar graph, and line graph just considered: what is the exact value of Variable 2 at level B of Variable 3 and level 4 of Variable 1? Most people find the question easiest to answer with reference to Table 6.1, a bit harder with reference to the bar graph, and hardest of all with reference to the line graph. This illustrates the principle of purpose-specificity, developed earlier and frequently noted in the graph comprehension literature, which is an inescapable consequence of the present theory: different types of graphs are not easier or more difficult across-the-board, but are easier or more difficult depending on the particular class of information that is to be extracted. In this case, we have already noted that absolute values of the dependent variable in a bar graph cannot be directly entered into the conceptual message, since there are no visual predicates that correspond to them. Rather, specific ratio values of the dependent variable can be encoded, as can pairings between arbitrary absolute values and ratio values (from the numbers printed along the ordinate); the absolute value of a particular entry must be computed by effortful inferential processes using these two kinds of information. The line graph is harder

---

[8] Bertin might motivate a similar prediction by saying that orientation is a retinal variable, and thus, according to his theory, may be apprehended in a single glance in a line graph, as opposed to the multiple glances necessary to detect the several heights indicating a trend in a bar graph (recall that his difficulty metric is the number of glances necessary to extract a piece of information). However, as we have seen, many predicates other than orientation may be used to convey trends in a line graph (e.g., "undulating"), and these are not to be found in Bertin's list of the 6 retinal variables.

[9] Incidentally, though a line graph is better than other forms of data presentation for illustrating trends, typically, only one way of constructing the line graph will illustrate a given trend optimally. For example, a line graph that used Variable 3 (i.e., A vs. B) as the abscissa, and Variable 1 as the parameter, would ot illustrate the linear and accelerating trends as transparently as the graph in Figure 21(b), since these trends no longer correspond to single attributes of a distinct perceptual entity, but must be inferred from the successive intervals separating the left endpoints of the five lines, and those separating the right end points of those lines, respectively.

still, because the Gestalt principles cause each entire line to be encoded as a single node rather than being broken up into a set of nodes, each corresponding to a level of Variable 1. Thus, when the conceptual question addresses the absolute value of Variable 2 corresponding to a particular value of Variable 1, there is no visual description node specific to the part of the line signifying that value, and one must be created by a top-down encoding process focused on a perceptually arbitrary point along the line. That is also why it is sometimes easier to use a bar graph than a line graph to determine the difference between two levels of one variable corresponding to a pair of values on another [e.g., whether the Consumer Price Index is higher for March or June in Figures 6.22a and b].

INSERT FIGURE 6.22 HERE

In sum, we have seen that extracting information from a graph is easiest when the visual description contains predicates linked to message flags displaying equations that answer the conceptual question. As a consequence, a) line graphs are best for illustrating trends and interactions (since there exist many visual predicates for line shapes); b) tables are best for illustrating absolute values of the dependent variable (since there is no way to specify these for particular levels of the independent variable in line or bar graph visual descriptions and graph schemas), and c) bar graphs are better than line graphs or tables for illustrating differences between dependent variable values corresponding to specific independent variable values (since the desired values are specified individually in the bar but not the line graph, and since it seems to be easier to encode a bar's height than to read a multi-digit number). Though the empirical literature on graph comprehension contains many methodological flaws (see Chapters 1 and 4), it is extremely comforting to know that these three conclusions have been borne out many times in that literature (Washburne, 1927; Culbertson & Powers, 1959; Schutz, 1961a, b; Carter, 1947).

Some Further Determinants of Graph Difficulty

In general, a graph maker will do best if he or she designs the graph so that the visual system parses it into units whose attributes correspond to the quantitative information that he or she wishes to communicate. In the previous secton, we saw how this principle favors either line graphs, bar graphs, or tables, depending on the type of question the reader is to answer. Of course, these are not the only choices that face a graph designer. In this section, we briefly show how other design choices can be resolved by the Graph Difficulty Principle.

1) One graph with two lines, or two graphs with one line? As mentioned, the visual system has predicates describing groups of nearby lines (e.g., Parallel (x,y), Fan-shaped (x,y,z), Intersecting (x,y), etc.). These correspond to specific types of interactions between variables (e.g., additive, multiplicative, inversely multiplicative, etc.). Thus, questions about interactions can be answered quickly if the lines are in close enough proximity for the predicate describing them as a group to be encoded. However, if the lines are in different graphs, they will be encoded as units, and their interactions must be extracted by interrogating their slopes separately and inferring the interaction from those slopes. Thus, when interactions are of interest, lines should be plotted on a single graph (unless, of course, the number of overlapping lines is large, which may lead to spurious groupings of line segments belonging to different lines). Schutz (1961 b) indeed found that graphs with multiple lines were easier to understand than multiple graphs, if the number of lines is small.

2) Legends or labeled lines? As noted earlier, the visual system groups stimuli that are in close proximity. A graph schema can exploit this fact by specifying that a label near a graph element signals the absolute value of a

variable for the conceptual message entry specified by that element (eg. in the bar graph schema we examined previously). If the correspondence is specified instead in an insert or legend (i.e., with a label next to a small patch sharing the color, shading, or internal structure of the lines or bars), that correspondence must be extracted by the effortful inferential processes, using one entry specifying the distinguishing feature of the bar or line in the graph, and a second entry linking that distinguishing feature to the appropriate absolute value, based on the legend or insert. Therefore, labeled lines should be better (again, assuming the number of elements is not so large that spurious groupings arise).

3) Grid lines or not? Whether a graph should include horizontal or vertical lines, aligned with absolute values on the axes and running across the graph, will depend on whether absolute values must be communicated (in a situation where they cannot be communicated by direct labeling, for example, the dependent variable in a bar graph). If absolute values are important, they can help in the following way: the top of a bar (or a well- defined segment of a line) can be perceptually grouped with a horizontal grid line, and the grid line can be grouped with an absolute value label on the ordinate, causing the nodes representing those elements to be linked in the visual description. The graph schema can attach a message flag to this node configuration, specifying the (approximate) absolute value of the dependent variable, and contributing a single conceptual message entry when the graph is read. Without the grid lines, as mentioned, inferential processes would have to deduce the absolute value by examining two distinct entries, one of which (the height of the ordinate label) would probably have to be extracted via top-down processes. Conversely, if absolute values are not part of the intended message of the graph, it is possible that the lines will form spurious groupings with graph elements, or may overload the capacity of the activated visual description, and in that case they are best avoided.

XI. The Efficiency of a Graph Reader

Until now, we have been referring to a single idealized "graph reader". Naturally, flesh-and-blood graph readers will differ from one another in significant ways. For example, some people may have swifter elementary information processes, or a larger short-term memory capacity, or more powerful inferential processes. Though these factors may spell extreme differences in how easily different people comprehend graphs, they are not specific to graph comprehension, and we will not discuss them further. Instead, we will focus on possible differences among people in their abilities to read graphs per se.

A natural way of determining what makes a person good at reading graphs is to examine what makes the graph reading process more or less easy (i.e., the considerations in the preceeding section) and to predict that individual differences in the nature of the structures and processes involved will spell differences in the general ease with which individuals read graphs.

Recall that in the last section, we showed that a given type of information was easy to extract from a given type of graph if there were message flags in the graph schema specific to that information, and if the predicates to which the flag was attached were present in the activated visual description of the graph. Each factor allows for individual differences. First a person's graph schema may lack important message flags. Thus, he or she may not know that parallel lines in a line graph signal the additivity of the effects of two variables on a third. When pressed to determine whether additivity holds in a certain graph, such a person would have to resort to costly inferential processes operating on a set of entries for ratio or absolute values. In general, the theory predicts that the presence or absence of message flags in a person's schema will have dramatic effects on how easily that person can extract the information specified by the flag. Second, the predicates that trigger the process whereby message flags are assembled into conceptual message entries may be more or less likely to appear in the visual descriptions of different people. The needed predicates, because of lack of practise at encoding them, may not yet be automatic, and hence may have low default encoding likelihoods. Furthermore, the links between those predicates and the rest of the graph schema may be weak, dissipating the "priming effect" which assists the encoding of missing predicates once a graph has been recognized.

Returning now to the first factor affecting the efficiency of graph readers, we might ask what will determine whether people have the necessary equation flags in their schemas, and whether the encoding likelihoods and links among predicates in a schema will be sufficiently strong. As to the first question, there are probably three routes to enriching graph schemas with useful flags:

a) Being told. It is common for formal instruction in mathematics and science to spell out what to look for in a graph when faced with a particular question. For example, students learning statistical procedures like the Analysis of Variance are usually told that nonflat lines indicate main effects, nonparallel sets of lines indicate interactions, U-shaped lines indicate quadratic trends, and so on.

b) Induction. An insightful reader or graph maker might notice that quantitative trends of a given sort always come out as graphs with particular visual attributes (e.g. quadratic functions yield U-shaped lines). He or she could then append the message flag expressing the trend to the predicate symbolizing the visual attribute in the graph schema.

c) Deduction. Still more insightful readers could infer that owing to the nature of the mapping between quantitative scales and visual dimensions in a given type of graph, a certain quantitative trend must translate to a certain visual property. For example, a person could realize that the successive doublings of a variable by an exponential function must lead to a curve that becomes increasingly steep from left to right.

Taken together, these principles suggest that improvements in the ability to read graphs of a given sort will come a) with explicit instruction concerning the equivalences holding between quantitative trends and visual attributes (so as to enrich the graph schema); b) with instruction as how to "see" the graph (i.e., how to parse it perceptually into the right units, yielding the appropriate visual description), and with practice at doing so (making the encoding process automatic and thereby increasing the encoding likelihoods and associative strengths of the relevant visual predicates); and c) with experience at physically plotting different quantitative relationships on graph paper (affording opportunities for the induction and deduction of further correspondences between visual attributes and quantitative trends, to be added as message flags to the graph schema).

XII. Extension of the Theory to Charts and Diagrams

Quantitative information is not the only kind that is transmitted by visual displays, and it would be surprising if the charts and diagrams used to express qualitative information were comprehended according to principles radically different from those governing graph comprehension. In fact, the theory described in these pages can be extended virtually intact to the domain of charts and diagrams. Again, a visual description of the diagram would be encoded, obeying the principles of grouping, the indispensibility of space, and so on, and again, there would be a "chart schema" for a particular subspecies, which specified a) the constituents of the visual description that identify the graph as being of the appropriate sort (e.g., a flowchart vs. a Venn diagram), and b) the correspondences between visual predicates and conceptual message entries. The conceptual message entries would be of a form appropriate to the qualitative information represented, and conceptual questions would consist of conceptual message entries with the "?" symbol replacing one of the constants. The MATCH, message assembly, interrogation, and inferential processes would play the same roles as before. Charts would be easier or more difficult depending on whether the visual system encoded them into units corresponding to important chunks of conceptual information, and chart readers would be more fluent to the extent that their chart schemas specified useful correspondences between conceptual information and visual attributes, and to the extent that those visual attributes were encoded reliably. A brief example follows.

Venn diagrams, used in set theory, consist of interlocking circles, each of which represents a mathematical set. Presumably, they are effective because the visual system can easily encode patterns of overlap (which will translate into set intersection), inclusion (translating into the subset-superset relation), nonoverlap (translating into disjointness), and so on. Simplified Visual Array, Visual Description, Chart Schema, and Conceptual Message representations specific to Venn diagrams appear in Figure 6.23a through d.

INSERT FIGURE 6.23 HERE

Even from these simplified examples, one can see that, as before, the difficulty of retrieving a given type of information will depend on what is in the visual description and graph schema, and not simply what is on the page. For example, here the reader would have to infer the fact that Set C is a subset of Set B from the conceptual message entry stating that Set B is a superset of Set C. A more efficient diagram reader might have a richer schema, containing the predicate "included-in" together with a message flag stating that one set is a subset of the other. This would spare that reader from having to rely on inferential processes.

Other sorts of diagrams and charts use other visual predicates to convey their messages efficiently: for example, flowcharts use shape predicates to signify the type of operation (e.g., action vs. test), they use the contiguity of shapes with lines to indicate the flow of control, and they use the orientation of arrowheads to indicate the direction of that flow. The linguist's tree diagrams for the phrase structure of sentences use horizontal position to signify precedence relations among constituents, proximity to common line segments to signify dominance (inclusion) relations, and above/below predicates to signify the direction of the dominance relations. For each type of diagram, there would be a specific schema spelling out the correspondence between visual predicates and conceptual messages.

XIII. Conclusions

This chapter began with a warning that our understanding of graph comprehension would advance in proportion to our degree of understanding of general perceptual and cognitive faculties. As we have seen, the theory outlined here indeed borrows shamelessly from perceptual and cognitive theory, adopting, among others, the following assumptions: the necessity of propositional or structural descriptions; the indispensibility of space as it relates to visual predicates, selective attention, creation of perceptual units, and accuracy of encoding; the limited capacity of short-term visual representations; the use of distributed coordinate systems for encoding shape and position; the perceptual integrality of certain physical dimensions; the use of schemas to mediate between perception and memory; the effects of physical salience on encoding likelihood; conceptually-driven or top-down encoding of visual attributes; a MATCH process for recognition; "priming" of visual predicates; and strengthening of associative links with practice. We trust that this enterprise is not totally parasitic, though, since in developing the theory, significant gaps in our knowledge of visual cognition came to light. For example:

- What are the exact constraints on the physical attributes that can serve as visual predicates, and what determines the likelihood of their being encoded?

- What are the relative strengths of the gestalt principles, and in what format should the groupings they impose be represented in structural descriptions?
- Which constraints determine how message flags can be appended to predicates in schemas? Are there limits on the types of predicates, the number of predicates, the number of parameters, and so on, that a message can refer to?
- How do visual descriptions guide top-down encoding processes?
- How general can the information in a general schema (like the general graph schema) be? Can such schemas be taught or enriched?
- What are the decay rates for different sorts of information in the visual description?

We would submit questions like these as particularly important targets for future research in visual cognition, ones whose answers will, in large part, be prerequisites for our further understanding of graph comprehension.

Finally, even in its current early stage of development the theory serves a useful role as a guide for constructing charts and graphs. In the following chapter we make use of it in our attempt to specify a set of complete guidelines for the creation of unambiguous, easily-understood charts and graphs.

CHAPTER 7: GENERATING CHARTS AND GRAPHS

In this book, we have focused on how people read and understand charts and graphs. We have approached this problem in two ways. We first considered the chart and graph as a complex set of symbols that work together to represent specific information. In Chapter 2, we developed an analytic scheme that specified how these symbols work, and allowed us to diagnose the reasons why they sometimes fail to work. A key component of our scheme was the set of "operating principles," most of which were rooted in observed facts about the operation of the human visual system, limitations on memory, and the way we comprehend symbols. These facts were pulled together in the theory presented in Chapter 6, which reflects the other side of our approach to the problem. In this case, we did not treat the chart or graph as a symbol system in its own right, but rather considered what would have to go on in a reader's head in order for that person to understand the information in the display. The theory provides an account about why the operating principles we have posited are as they are, and about what underlying factors result in us needing a display with certain properties.

The theory just presented serves an important role when one wants to draw a chart or graph. In this case, from the outset one wants to avoid violating the principles we have posited; it is awkward to draw a display first, and then analyze it, and then repair the flaws. Obviously, it is much better to keep the potential problems in mind from the start and simply avoid succumbing to them. One way to do this is to try to put oneself in the head of the reader. First, try to specify exactly what information you want the reader to come away with when he or she reads the display, and then consider how best to ensure that that information gets there. Thinking about things in this general way will lead you first to specify your message, then to select the graph type (which is equivalent to selecting a combination of a type of framework and specifier) that will be the best vehicle for that message. (See also Wright, in press, who emphasizes that this "psychological approach" helps a designer formulate better documents and computer readouts.) Once having selected a type of framework and type of specifier, it is relatively straightforward to use them to represent particular information effectively. In so doing, one must also keep in mind that the pragmatic factors described in Chapter 4 can add or shift.

In this chapter, we offer a step-by-step guide to generating effective charts and graphs. This procedure is based on the analytic scheme we developed and tested earlier. Thus, the procedure leads one to construct displays in terms of their basic constituents and to do so within the maxims of the various operating principles detailed in the previous chapters. Our guidelines are sufficiently precise that we believe they can be developed to the point of being incorporated in a computer program (and are now working on doing just this). However, this program would have to interact with human users because there are some questions that only the user -- who knows the context in which the display will occur -- can answer. Furthermore, often the user does not entertain these crucial questions on his or her own, but must be prompted to answer them explicitly. These questions must, for the most part, be resolved before one actually lays pen to paper (or pushes keys on a terminal), and thus we consider them below before turning to the nuts and bolts of constructing the actual display.

As was evident in the previous chapter, charts and graphs convey information at different levels of precision. A rising line is a kind of graph, although all it conveys is that something is increasing relative to something else. In such a display, the implied framework is nothing more than an assignment of direction, indicating which way the relevant values are increasing. In this chapter we give instructions about how to generate the most demanding, precise kind of display. If the purpose at hand does not require such precision, the superflous added information should be deleted. It is up to the illustrator, however, to decide exactly what information is relevant and what is not; if this is decided from the outset, the system can be used as described, only now certain parts of our advice will be superfluous. For example, if the illustrator decides that the actual values of variables are irrelevant, he or she may simply ignore all advice about labeling axes and ensuring accurate reading of specific points. Thus, before beginning we must have a clean idea of what we wish to accomplish.

1. The initial analysis

Before one can begin to draw a display, one must first answer five questions: 1) What information should be in the display? 2) What is the purpose of the display? 3) What impression do you want to convey? 4) Who are the intended readers? and 5) What materials do you have to work with? Let us consider each question in turn and consider the sorts of factors that will enter into your decisions.

What information should be in the display?

Deciding on what you want the reader to know after reading a given chart or graph is critical. Before doing anything else, you must decide what information you want to convey. A useful heuristic here is to think of a title for the display. For example, "Change in productivity over time" would lead to a different display than "Amount of oil produced." In the former case we would certainly plot values over time, whereas in the latter we might choose to dispense with time and present output from different countries collapsed over time. If time in fact is irrelevant to the intended message, dispensing with it might save the reader effort and possible confusions. Only after this question is resolved can one know which data are relevant.

What is the purpose of the display?

The second question follows naturally from the first. Given one has decided on the data to be presented, what level of detail is necessary? Should the reader use the display to extract specific measurements or just to get an overall impression? If too many data are present, they will have to be boiled down into a relatively small number of averages (as will be discussed shortly), and it is up to the graph maker to decide which levels of detail must be sacrificed. Is it necessary to know data about every day of production, for example, or are monthly data satisfactory? If the reader is to extract arbitrary levels of detail, and one has hundreds of numbers, a graphic display may not be appropriate at all to this "archival" function.

What impression do you want to convey?

At this point, you should have a set of numbers that could be displayed using several different graph types (as will be discussed shortly). Before selecting a graph type (i.e., type of framework and specifier), you should decide two more things: Do you want to emphasize or de-emphasize a given relation? If so, what is it? If you wish to emphasize the rate of growth of one variable over another, you should keep this in mind when selecting a framework. Recall that various pragmatic factors will vary the impression a chart or graph conveys. If need be, flip back to Chapter 4 and briefly review these principles. Keeping them in mind, we will soon see that different frameworks are more or less pliable for use in exploiting specific principles. In addition, if it is necessary to decorate a page as well as convey some data, this should be kept in mind when selecting among the range of possible frameworks and specifiers. Artists often use depiction so that the framework, specifier and/or background reinforces the basic message. For example, to present information about unemployment, they are tempted to use a line of people waiting for unemployment benefits as bars in a bar graph. Tempting though such decorations are, however, we stress that it is far more important not to violate any of the operating principles. In the last section of this chapter we present a step by step procedure that should prevent an artist from doing so, even when the display is quite unorthodox. In the fifty cases we have tried, this scheme proved adequate, and we intentionally varied the kind of depictions and techniques used in an effort to strain the system. Further, the system is so explicit that we have written a computer program for the APPLE computer that produces violation-free charts and graphs (write the authors for more information about the program). Thus, we are confident in recommending our procedure for use in designing most displays.

What is the intended audience?

The nature of the intended audience is important for two reasons: First, the concepts you explicitly label in the graph obviously must be familiar to the intended readership. For example, plotting first or second derivatives and labelling the axis as such excludes people who have not studied calculus. The same information could be presented by plotting the simple level of a variable, and allowing the first and second derivative to be read from the graph as slope and curvature respectively, which is a simpler and more accessible concept. In addition, no exotic words should be used in labels, nor should uncommon symbols be employed. Second, the graph type used is to some extent dependent on the readership. The graph types we will discuss here are all common to most literate people, but there are others that are less common, with visual patterns that are not obviously translatable into quantitative trends. For example, in engineering studies there are diagrams in which information is displayed as blotches whose shape represents information in a polar coordinate space. These graphs are quite interpretable to one thoroughly familiar with them, but are only a hindrance to the rest. In many cases even the common graph types we will discuss may not be universally known, in which case one has little choice but to use tabular presentations. (In fact, for simple and small data sets, tables are comprehended quite well, even by children; Wainer, 19XX; Wainer & , 19XX).

What is there to work with?

The final thing to keep in mind is rather basic: What physical materials are there to work with? Can color be used? Is the display going to be in a small area of a page, or on a large bulletin board? Will the display be on a computer graphics screen with a coarse grain? Can you vary the weight of lines? The size of letters?

These five general background factors must be kept in mind before one begins construction of a display. Only after resolving these questions can one intelligently proceed to the next step, deciding on the structure of the display.

2. Choosing the correct display type

Having decided what data one wants to illustrate, what use the display will be put to, what overall impression is to be conveyed to which readers, and what materials and on hand, one is now in a position to begin drawing (drawing taken in its broadest sense, to include displays on a CRT).

Charts vs. Graphs

The first question that must be asked is, what is being related to what? That is, are the relationships you wish to convey essentially qualitative or quantitative? Recall that in Chapter 2 we distinguished between these two kinds of semantics, noting that charts usually convey information about qualitative relationships (such as "is a member of" or "occurs after") whereas graphs always convey information about quantitative information ("x has more than y"). We pointed out that there are a number of kinds of relationships possible in both cases.

To review, when viewed as format symbols, both charts and graphs convey information by relating parts of a framework together. Charts do so by connecting distinct framework elements (usually boxes or nodes) with arrows or lines. The relationships symbolized can be directed or nondirected. A directed link is not symmetrical: for example, an organizational chart has links labeled "under supervision of" or the like which point down. A symmetrical link, such as "sibling of" in a family tree, is equally valid going either way. The relationships in charts can also be all of a single kind (as in a flowchart, where all links mean "followed by") or can be of multiple kinds. If they are of multiple kinds (such as would occur in a family tree), the different kinds of links must be clearly distinguished and labeled. Finally, a given part of the framework can be related to one other part or to many other parts, depending on what is being discussed. For example, one-many mappings characterize hierarchical structures, and one-one mappings characterize flow charts.

If the kind of data you have is of this general type, where distinct entities are being related qualitatively, then you want to draw a chart. In drawing the chart, first decide on the basic structure (hierarchical tree, sequential steps, etc.). Then consider the steps discussed in the third section of this chapter. Be sure that the important relationships to be conveyed take the form of easily perceived visual patterns, as will be discussed shortly.

Graphs represent information by pairing an extent associated with one axis with a position or change of position along the other. In this case, the specifier serves as a function, with each relevant point along it pairing a point or region on the horizontal axis with a point or region on the vertical axis. Graphs relate two different scales together, and depending on what kind of scales are being related, different graph types are more or less appropriate.

To review briefly, there are five scale types. Nominal scales are not ordered at all; numbers or other symbols are used as labels (as with company names, numbers on athlete's sweaters or the like). Ordinal scales are rank ordered only; the actual magnitudes of differences are not reflected in the ordering itself (the difference between first and second may be twice as great as the differences between second and third, but this will not be evident in this kind of scale). Interval scales preserve the actual quantitative differences between values (such as farenheit degrees), but do not have a natural zero point. Thus, ratios cannot be taken among items on an interval scale; 10 degrees farenheit is not twice as cold as 20 degrees farenheit. Ratio scales are like interval ones but they do have a natural zero. Thus, not only do quantitative differences among values have meaning, but so do ratios. Two hundred dollars is twice as much as 100 dollars. A fifth scale type, absolute scales, are ratio scales with non-arbitrary units: number of jellybeans in a jar for example, unlike dollars, which could be changed to different units (e.g., cents) with no loss of information.

Thus, if you are relating variations in some quantity to something else, you want to use a graph. In many cases, one of the things being considered is a set of names (i.e., a nominal scale - company, country, condition, etc.); the most frequent exception occurs when changes over time are considered, in which case a ratio scale (time) may substitute for a nominal scale. In special cases the other scales may serve the same purpose. In all these cases, one is comparing a number of things with respect to a single scale of measurement. When more than one scale of measurement is involved (i.e., several non-nominal scales are mapped onto a nominal scale), we recommend that the choice of separate frameworks versus a single framework be made according to the following criterion: if the similarities or differences among the non-nominal variables are part of the intended message, and if the number of such variables is not too large, then a single framework should be used. This allows overall similarity of differences or trends to be displayed as parallelism or various sorts of nonparallelism (e.g., fanning out), which can be perceived as entire slopes without the need to glance back and forth between graphs. However, in such cases one must be aware that when one uses a single framework to represent more than one measurement scale, it is difficult to signify how values are related to specific specifiers. In many of the cases we have seen, trying to include more than one scale for dependent measures in a simple graph results in ambiguities and an incomprehensible display (e.g., see the second example at the end of Chapter 2, where the middle framework was used as two scales). Thus it is important to use similarity of color or shading, or explicit labels for each line and scale, so that the correspondence is apparent. As a corollary, if the number of different scales is large, or if similarities or differences in trends are irrelevant to the intended message, then separate graphs should be used.

Choosing the correct chart or graph type

Charts. If one is dealing with a chart, the choice of a graph type is almost entirely dictated by the nature of the connections between the things represented. If one is dealing with one-many mappings, where each thing is connected to two or more others, and each of these in turn is connected to two or more other things, a hierarchical scheme is dictated. The convention is to put the [illegible]ments of the framework (boxes, nodes, depictions) such that the elemen[illegible] at the "dominating" end of their relationships with other elements ("dominating" meaning "supervising" or "including" or the like) are higher on the page. If you are dealing with one-one mappings, the nature of the specifiers dictates the framework again, with temporal sequencing requiring a left-to-right organization in this culture. The constraints on chart construction thus arise not so much from the general nature of the specifier and framework as from the operating principles, especially those that proscribe violating the limited processing capacities of human readers. These principles are incorporated in the specific instructions to follow.

Graphs. In the following sections we will consider when each of the five most common graph types is appropriate for each type of data. The graph types we consider all are in common use and are relatively general purpose. Let us first describe these alternative graph types:

A pie graph consists of a circular framework which is divided into a set of wedges. Each wedge represents a percentage of the whole, as indicated by its relative area.

A divided bar graph is like a pie graph, but the framework is rectangular. In addition to having the internal area divided vertically into a set of smaller rectangles, each above the other, in such graphs it is common to have a scale marked along the left side of the framework.

A line graph usually occurs in an "L" shaped framework (which is sometimes closed into a rectangle), with the scale of the dependent variable (i.e., thing measured) associated with the vertical axis and the scale of the independent variable (things that measures were taken of) associated with the horizontal axis. A line (or lines) serves as a specifier, providing specific values, differences, and trends of the dependent variable for specific values, pairs of values, and ranges of the independent variable. The height of the specifier line over a value of the independent variable corresponds to the value on the vertical scale at that height; the shape and slope of the line as a whole corresponds to the difference or trend of the dependent variable paired with the range of the independent variable that the trend being examined sits over. Lines are continuous, representing each point on the x axis.

In general, a bar graph is like a line graph except that bars usually stand over the labeled locations on the x axis. The height of the bar indicates that the value at that height on the vertical axis should be assigned to whatever is labeled under the bar. Bar graphs can be constructed with the bars being vertical or horizontal; when bars are horizontal, the scale of the dependent variable now is on the horizontal axis.

A surface graph is simply a bar graph in which the bars are so wide that they are connected, flowing horizontally into one another. As in a bar graph, the area within each bar is often shaded.

We have chosen not to treat one last common graph type as a distinct class. Pictograms are simply bar graphs in which the bar is replaced by a stack of identical pictures. Usually each picture represents some fixed number of the units of measurement (e.g., each barrel may stand for 1,000 barrels of oil produced). These graphs function just as do normal bar graphs, with the height of the bars indicating the value of the particular thing being measured; the number of pictures is completely redundant with height. The only cases in which this is not true involve the unit picture being assigned a value in a key and no vertical axis is included. It is conceivable that there are special circumstances in which this is a desirable feature, but it is not apparent what are the general principles that will identify such situations. Thus, given that depictions can also be used for all of the other graphic constituents, we did not consider this one case sufficiently different from standard bar graphs to warrant a separate category. Rather, pictograms result when depictions are used pragmatically to reinforce the point of the graph, or to convey absolute amounts by allowing the reader to count symbols.

Finally, we have not discussed location graphs, which usually consist of a map with different symbols over different locations (the symbols represent things like the population or temperature at the locations). These graphs are not general, but are used only to map values of a dependent variable to specific locations. They do not function as do the more general graph types, with each spatial dimension standing for a different nonspatial or conceptual dimension. Instead, each spatial location on the graph represents a spatial location in the world. Thus, these graphs are in fact misnamed: they are really simply maps, with particular information being supplied in addition to location (as may be done in road maps indicating height or population density by different colors). In this book, we have explicitly excluded discussion of maps, for the reasons discussed in Chapter 2.

Two major determinants of the best graph type to use are the nature of the scale of measurement (the "dependent variable") and the nature of the things being compared (the "independent variable"). But these are not the sole determinants. In addition, properties of our perceptual and memorial systems favor some graph types over others for specific purposes. Thus, we must consider two things when choosing a graph type: The nature of the data, and the purpose to which they will be put.

Let us now consider in more detail how different factors affect the choice of a graph type. First, the five types of scales can be further divided such that proportion and percentage are differentiated from other types of amount. If this is done, we have four classes of measures: Ranks, proportions, intervals, and ratio scales. We must consider the appropriateness of each framework type and specifier type for each kind of measure when the items arranged along the x scale are themselves ordered on a nominal, rank, interval or ratio scale. Thus, we have four possible measurement scale types and four possible independent variable scale types, resulting in 16 unique pairings. But this is not all there is to it, sad to say. We must also take into account the purpose of the display, which often will be the deciding factor when multiple options are technically appropriate. Recall that we know that people have a difficult time seeing a single part of a perceptual unit and comparing it to another unit or part thereof (see Chapter 3). Thus, line and surface graphs are to be avoided when specific point information is being conveyed. On the other hand, people have a limited capacity for apprehending information and making comparisons. Thus, bar graphs are to be avoided if numerous (more than four) points are presented, or if a number of pairwise or n-wise comparisons must be made. Line graphs, in contrast, can usually be encoded as composed of relatively few higher-order perceptual units (i.e., upward rising lines. "U" shapes, etc.), which minimizes problems due to memory capacity limits, and allows abstract patterns of differences or trends to be depicted as single visual properties. In addition, to considering which formats are least taxing for a specific purpose, the reader will have to consider which formats are easiest to modify to emphasize a particular point.

A. Rank data.

Rank data cannot be presented in a pie graph or a divided bar graph (which is really just a rectangular pie with rectangular, stacked slices). Of the remaining graph types, the one chosen will depend in part on which scale is used along the x axis, as noted below:

Nominal scale along the x axis: If items along the x axis have no specific ranking, bar graphs are in general the most appropriate graph types. If a bar graph is chosen, horizontal bars may be used if there is no inherent ordering among the measured things (although vertical bars may be preferable because of their familiarity alone); if the things are ordered in some way, a vertical format is preferable, with the bars being ordered left-to-right along the x axis. It is often a good idea to order entities from the greatest to smallest along one of the scales used. As XXX (19XX) has pointed out, usually the reader has an expectation as to the order of entities, so that mere presence of an item in an unexpected position itself conveys information. Furthermore discontinuities in the sequence (e.g., if the wealthiest 20 countries are far richer than the poorer 140), will be apparent in a large step at the point of discontinuity. When large discontinuities exist, and the artist wishes to emphasize them, a surface graph (i.e., a bar graph with no space between the bars) is best.

Under most circumstances, one might be tempted to eschew a line graph when the x axis represents a nominal scale, reasoning that a line graph confuses the reader by suggesting falsely that some continuous scale, with interpolable intermediate values, is represented by the x-axis. Many standards for graphs recommend this convention. However, there are two circumstances in which a line graph, surprisingly, is appropriate. First, if the number of x values is small, and there are several dependent measures that behave in different ways with respect to the members x, a multiple line graph will translate those differences into salient nonparallel lines. This allows the reader to detect the difference by encoding the nonparallelism directly, without having to examine the graph bar-by-bar (see Chapter 6). The second case where a line graph is mandated for nominal data occurs when a) the graph format is standardized, so the left-to-right order of nominal scale values along the x axis is non-arbitrary in the sense that it is the same for everyone, and b) various patterns of values of the y-scale with respect to the x must be differentiated. In these cases, a line graph allows each pattern to be represented as a line with a different, universally recognizable contour -- again sparing the reader from having to undertake a cognitively costly element-by-element comparison. One notable example is the Minnesota Multiphasic Personality Inventory, a personality questionnaire yielding nine values usually represented as a line graph called a "profile". In the lore of MMPI interpretation, a line sloping up to the left represents a neurotic personality, a line sloping up to the right represents a psychotic personality, and among the more sexist psychotherapists, a double-peaked "spread legs" profile represents a hysterical personality, and so on.

Ranks along the x axis: These kinds of data should be graphed using the foregoing guidelines, but with the following additional conditions:

In one class of cases, the ordinal scale is used simply as a convenient way of ordering essentially nominal values; for example, oil production of different countries arranged by decreasing size, where no particular relationship of production to size is important -- the ordering by size only helps the reader find a given country, or a discontinuity in the sequence, easily. The rules for nominal scales apply to these class of cases.

A related class involves truly irreducibly ordinal scales, such as rank in a class, or finishing order in a race. In these cases, a bar or surface graph is preferable to a line graph, all other things being equal, since the latter connotes some irrelevant or nonexistent x-continiuum against which trends in y may be usefully examined. However, as with nominal scales, ordinal scales may be represented with line graphs if the difference between two rank orderings is being compared, so that differences in rankings can be perceived as patterns of lines. It seems reasonable to assume that readers will be astute enough to read a line graph representing an ordinal scale and not change their conception of the scale as a result (e.g., it is unlikely that they will think that finishing order in a race is an interval scale just because they have seen it represented in a line graph!)

In another class of cases, ordinal data are used to sample or exemplify portions of what is conceptually an interval scale. For example, one might want to illustrate that large populations are associated with high infant mortality rates. In this case, countries would form an ordinal scale if ordered by population, but it is not the identity of the countries per se (as in a nominal scale) or the order per se (as in a race) that is conceptually important. It is the population scale itself that is conceptually an interval scale, and which must be invoked in order for a reader to conceive of the universal, lawlike relation being communicated (i.e., the precise countries are unimportant; hypothetical countries with intermediate values could be used to the same effect). In such cases, there would be no reason not to use line graphs, even though they are ordinarily most appropriate for interval or ratio data.

Interval or ratio scale along the x axis: The bar, surface, and line graph formats are all appropriate for these kinds of data. Note, however, that the baseline is arbitrary for interval data. Thus, axis truncation can be used freely, as convenience dictates. The main determinant of which graph type to use is the purpose to which the display will be put. If point information or the amount of difference is important, a bar format is more appropriate; if slope or trend information is important, or comparisons among slopes, differences, or trends, a line graph is more appropriate.

Percentage or proportion data

A. Single Nominal scale . When this scale is the independent variable, percentage or proportion data can be graphed using three different graph types. The most common format is a pie graph, with the relative area of slices representing the proportions of the represented quantities. This format has two important limitations, however. First, precise comparisons cannot be made in most cases because it is very difficult to measure on the framework appropriately without a protractor (though a series of tick marks arrayed around the circumference of the pie can help). Second, no more than four or five slices representing message-relevent entities should be used. Thus, if a number of different kinds of things (i.e., different specifiers for each independent variable) are compared, a multiple framework display or one of the following graph types will be required.

The second option is a divided bar graph. This format is a kind of square pie graph, with the length of each bar representing the proportion. A scale is commonly included but can be misleading if numbers we associated with it. The height of each bar does not represent the proportion, as one could mistakenly conclude in Figure 7.1.

INSERT FIGURE 7.1 HERE

The advantage of this format is that precise values can be represented (use of an inner framework grid can allow one to read specific values and compute lengths quite easily) and multiple different things can be represented in the same framework (by different bars). Thus, this format has about 4 times the information-capacity of a pie, given that people can process about four bars, each of which has as many internal divisions as can be used in a pie graph.

The third option with these data is a bar or surface graph, with percentages being represented along the vertical or horizontal axis. In this case, however, the reader will not see how the whole is divided up, with the quantities of the various entities necessarily in exact inverse proportion. All that can be seen are differences among parts.

B. Nominal scale plus another scale along x axis. Matters become more complex when proportions among a set of nominal values are then contrasted at several levels of an additional independent variable. Figure 7.1 shows the three different ways of graphing the same data when two independent variables are considered. Note that in the pie graph you can see relative amounts easily at a single level of the variable distinguishing the different pies (time, in this case), but it is difficult to compare actual amounts across the multiple frameworks. This kind of comparison is easier with the divided bar graphs, but the relations among the individual components are not as transparent. In the final case, where we have separate bars for the two kinds of soap, the trends in percentage of people using them are visible, but their status as proportions, with an increasing share of one entity necessarily eating into the share of another, has no direct counterpart in the visual description of the graph. Thus the multiple divided bar graphs seem best when proportions of nominal scale values are contrasted over a second nominal scale, or a truly ordinal scale (see page ___ above).

A close to ideal case for proportions varying over some interval or ratio scale is what we can call a "line/divided bar graph". Here the relative widths of segments that partitions the area of a rectangle horizontally can change continuously from right to left (see Figure Z). In this case, change in proportion over time translates into tapering left, tapering right, bulging, or saddle-shaped segments, and the reciprocity of the proportions of different segments at a given time is evident from the fact that all the segments must be squeezed into a rectangle of unchanging height. One potential problem with this format is that the slope of segments, though perceptually available, is conceptually irrelevant, and may even interfere because the slope is perceptually integral with the conceptually relevant dimension of segment width (see Chapters 3 and 6). One way to counteract this problem is to stack the segments in a top-to-bottom order that reduces overall slopes and emphasizes widths -- thus Figure X(a) is a superior to the same information graphed in a different order in X(b).

4. Selecting the axes

At this point, the reader should be able to select a graph type for a set of data. The question now becomes which of a number of independent variables should be placed on the x axis. That is, in many graphs there are multiple specifiers. Each specifier is labeled, and in fact the entire set of these labels could just as easily have been placed on the x axis, with the labels originally layed out along the x axis now being paired with individual specifiers. Figure 7.2 illustrates such a case. Once again, the choice of layout depends on the purpose of the display. There are four rules of thumb here:

INSERT FIGURE 7.2 HERE

First, the designer should decide which independent variable (times, conditions, years, etc.) is composed of entities are to be contrasted with one another (let us call this the "foreground variable") and which variable or entities are to serve as a backdrop for the comparison of interest, "backdrop" in the sense of serving as a set of occasions for the contrast within the first independent variable to be made repeatedly. For example, in Figure 7.2a, the information is naturally summarized as "in 1960 the U.S. was much more productive than Japan, whereas in 1980 the U.S. was only somewhat more productive." Here the foreground contrast is between countries, and it is made twice, against the backdrop of different years. However, 7.2b, containing the same information, is summarized as "U.S. productivity declined between 1960 and 1980, whereas Japan's increased", with years as the foreground contrast, and countries as the backdrop. The rule of thumb seems to be that the foreground variable should be drawn as the parameter (with specifiers labeled by the individual values), and the background variable as the abscissa.

Second, the designer should consider which contrast in trends is to be emphasized, which may not be the same thing as deciding which variable is foreground and which variable is background. If a clear contrast is intended, the variables should be assigned to specifiers so that the relevant contrast appears as a recognizable shape. For example, 7.2a seems far easier to read than 7.2b, perhaps because the narrowing of the productivity difference, the graph's intended message here, comes through as a converging pair of lines, whereas in 7.2b it comes through as a difference in slopes, which does not connote a "narrowing" of differences as saliently.

Our third rule of thumb, is straightforward: an interval scale is highly suited to a continuous axis, whereas countries, a nominal scale, is more suited to a set of discrete specifiers.

Thus, all other things being equal, if finally, when one independent variable has a smaller number of levels than the other, the smaller should comprise the parameter (labeling individual specifiers) and the larger the abscissa, so as to reduce the number of visual units in the graph, and to allow the complex comparisons within the multi-valued variable to translate into the shape of a line rather than the differences among a set of endpoints (compare Figures 7.3a and b.

INSERT FIGURE 7.3 HERE

These four heuristics can often conflict if applied outside any particular context. However, these conflicts can usually be resolved if one considers the point one is trying to convey and the use to which the information will be put. If these factors cannot resolve a conflict, it is likely that any arbitrary resolution will be satisfactory.

5. One framework or many?

The final general question that must be settled before a display is generated is straightforward: Should the information be presented in a single framework or in multiple frameworks? The main factor governing this decision is complexity: There should not be too many perceptual units within a single framework. Thus, if there are more than four functions being plotted in a line graph, or one bar per point in a bar graph, or five slices in a pie chart, or one surface in a surface graph, then a multiple framework may be appropriate. One exception to this rule occurs when the Gestalt laws group sets of lines into higher-order units: when all or almost all the functions are similar and the point of the display is to emphasize this fact, all the lines should be plotted in a single framework. Similarly, when the functions fall into two groups, with the functions within each group being similar to one another, a single framework may be used.

When multiple frameworks are used, the designer must decide what will go in each framework. The selection of which independent variables to put in which frameworks should be governed by their similarity and their relevance for each other. That is, similar categories should be put in the same framework (if only to make it easier to comprehend what is there) and variables that will be compared together should be placed in close physical proximity so as to constitute higher-order shape patterns which can be perceived as units, this signifying a trend or pattern of trends directly.

Finally, multiple frameworks often seem appropriate when several different dependent variables, measured in different kinds of units, are related to the same independent variable and are meant to be compared with each other. The reason that multiple frameworks are generally appropriate is that the different sets of scale units along the y-axis are not easily linked with the appropriate specifiers within the graph. There is one have obvious exception to this rule: when the number of specifiers is two, and they can be clearly related to two vertical axes in a U-shaped framework by arrows or perceptual similarity (e.g., when the lines are the same color and boldness as their respective axes), then placing them on the same graph yields the advantages mentioned earlier: differences in the respective trends can be seen easily, as the nonparallel shape formed by the two lines.

When two or more frameworks are used to display dependent variables measured in different scale units, the measurements are in fact scaled arbitrarily with respect to each other, and thus the different frameworks need not be the same size. For example, one may want to compare number of suicides per year with the rising cost of food. Making both frameworks the same size allows for easy comparisons, but making one bigger emphasizes the point that suicides have been rising, as is illustrated in Figure 7.4.

INSERT FIGURE 7.4 HERE

3. Guidelines to drawing

The following guidelines should be used when drawing a chart or a graph. However, many of the specific pointers are only relevant to graphs, which can simply be ignored when one is drawing a chart. In addition, we often use terminology specific to graphs per se, as was done in Chapter 5; the reader should realize that "axes" refer to a part of the framework, and usually correspond to "box" or "node" if one is drawing a chart instead.

Drawing a multiple framework display

When laying out the separate subgraphs in a multiple framework display, the following guidelines should be obeyed. If you don't need a multiple framework display, skip this section.

1. The physical arrangement should lead the reader to examine the displays in a logical sequence. Readers in our culture will examine displays left to right and top to bottom. If a particular order is critically important, subdisplays can literally be connected by arrows indicating order-of-inspection.

2. The relative visual salience (reflected by differences in line weight, color, and size) should reflect the relative importance of the information presented in each display. Make the more important sub-display bigger or in some other way visually striking; if no display is more important, make them all equal size and equally salient.

3. The individual subgraphs should be clearly labeled. The labels should be closest to the appropriate subgraph, such that they are clearly associated with the correct display.

4. If the same specifier elements are used in more than two subgraphs, use a legend to supplement direct labelling. Make sure that labels in the legend are clearly associated with the appropriate specifier elements. In this case, pair each label with a small segment of the relevant specifier element. Make sure the specifier elements are highly discriminable. Do not have more than 4 labels in a single legend. (If you need more, be sure to label some of the specifiers directly, even though there may be redundancy.)

5. If the same variable is discussed in two or more subgraphs, make sure the subgraphs have the same general form, with the variables being presented in corresponding locations on the framework (for example, two pie charts of the same data at different times should have data presented in corresponding "slices").

6. If the same variable is discussed in two or more subgraphs, the same units ought to be used along the framework. Unless there are extreme range differences (e.g., orders of magnitude), the units ought to be laid out using the same number of ticks per centimeter, starting at the same baseline.

7. If one subgraph presents a second version of the same information presented in another subgraph, this should be clearly specified in the titles or by arrows showing the correspondence. If arrows or other visual means are used to establish the correspondence (e.g., a drawing of a magnifying glass), you must be sure that it is clear how one subgraph relates to another, even if

additional labels are needed to specify the relationship (e.g., "When the years 1980 and 1981 are examined in detail").

8. The relationships between the subgraphs should be unambiguously specified by titles and/or by other visual means associated with the subgraphs by the Gestalt laws.

9. To the extent that subgraphs perform different functions, they should look sufficiently different so the reader will not assume they are showing the same thing (e.g., the scales should be labeled using different types of font).

After deciding on the graph type for each of the subgraphs in a multiple display and deciding how they will be organized on the page, each individual display should be drawn according to the following guidelines. In this case, however, one should also keep in mind the general guidelines just provided, varying font or keeping it constant as is desired to emphasize differences and similarities in the information conveyed in the different subdisplays.

Drawing the Framework

The outer framework is the first thing to draw. When doing so, keep the following rules in mind:

1. The marks that define the outer framework must be grouped together by the Gestalt principles so that the framework is clearly defined. Every necessary part must be present or obviously implied.

2. If tick marks are used between scale values, there should be no more than five before a heavier tick mark or a new scale value.

3. The marks must be congruent with the idea being conveyed. Thus, an ordinal or nominal scale must be clearly demarcated.

4. If the framework depicts, the depiction must be representative of the class of things it stands for. Further, the depiction must be chosen and drawn so as to be unambiguous.

5. Ideally, all parts of the framework should play a role in communicating quantitative information. If for some reason (e.g., you use a particular depiction) they do not, make the superfluous parts lighter than the rest of the framework or clearly set aside.

6. The axes should be uniform and continuous; if they are not, be sure you are distorting things to make a particular point, in a way that the reader can detect and understand.

If readers are expected to extract precise information, an inner framework is useful. The inner framework should be chosen after the specifier elements are in place. This is because you do not want the placement of inner framework elements to group with the specifier elements; choosing the inner framework after the specifier is in place allows one to avoid this pitfall. We will thus discuss construction of the inner framework after discussing construction of the specifier.

Labeling the Framework

Before you can put in the specifier, you need to label the framework. This is critical in a graph where you need to know what each axis represents and how the scale is constructed on each axis.

1. Put on a title. The title should state clearly what is being graphed or represented. The title should be recognizable as such because it is clearly set off from the rest of the graph; it should not be close enough to any line to be grouped perceptually with it. A larger font size will also prevent the title from being perceptually grouped with other labels or parts.

2. Label each axis. The labels should be placed closer to the axis they label than to anything else, ensuring that they will be grouped perceptually with the right axis. Labels parallel to their axes are a good choice in complex displays because the Gestalt law of common fate will group label and axis together.

3. Put scale values on both scales; make sure they are closer to the correct tick mark than to anything else.

4. Make sure all labels are legible and will remain legible if the figure is reduced. (Chapter 3 presents a way of computing this beforehand.) Use arabic numbers, not roman numerals, and avoid italics (Wright, in press).

5. If you use depictions as labels, make sure the pictures clearly stand for what you want to label. A quick way to test this is to ask several people to provide the first name that comes to mind when they see the picture; this name should be the label you have in mind.

6. Use words that are consistent with the text in which the display will be embedded and with common usage about the topic.

7. Keep the graph as close to the text as possible.

Drawing the Specifier

1. Make sure that the specifier elements are easily seen. Relevant differences in values must be discriminable even after photoreduction has taken place.

2. Make sure different specifiers are clearly discriminable. Different shading should be used with different bars, pie-wedges, and surfaces (but see below), and different widths and patterns should be used with lines (but see below). If more than one line is present, make sure all the segments of each line clearly are grouped together; this requires having different lines drawn in different widths, patterns and/or colors, such that the segments of any given line are more similar to each other than to anything else in the display.

3. In order not to mislead the viewer, do not vary irrelevant integral dimensions (e.g., height and width of a bar).

4. If shading is used, make sure differences in shading line up with the values being represented. The lightest ("unfilled") regions represent "less," and darkest ("most filled") regions represent "more". Similarly, in a divided

bar graph, shading of bars should proceed from lightest (unfilled) to darkest (filled) going from top to bottom.

5. Avoid unnecessary depictions incorporating the specifier. If such decoration is irresistable to you, make sure it is representative of the point of the display. Also, make sure that the role of the lines as a specifier is not lost in their role as a picture.

6. If specifier elements abut, make sure it doesn't look like they overlapping; have a sharp line between them. If they overlap, make some of the lower one protrude from under to top one.

7. Make sure there is a visible change in the specifier element every time it represents something different. If a single line is used over a period of trials until a treatment is added, for example, make sure this point is marked somehow. Every meaningful difference should be clearly indicated by a perceptible difference in the marks, and vice-versa.

8. If color is used, be sure that the most important specifier element stands out the most; if no one element is more important, avoid using hues of different intensities or saturations for the different elements.

9. If color is used, do not use values from the entire color scale to represent quantitative values (colors don't fall perceptually along a single continuum). If colors are used as a scale, use variations of saturation, or if necessary, variations within the red-orange-yellow (in that order for low to high) family or the green-blue-violet family (as ordered); these variations do somewhat fall into a continuum perceptually.

10. If color is used, make sure that values of color do not contradict cultural conventions (red is hot, green for safe, etc.).

11. If 3-D perspective is used, remember that volumes and areas are not accurately read; avoid perspective effects if you want to convey precise values. Also avoid sharply oblique viewing perspectives, which distort quanti-

ties, or extra lines that turn 2-D surfaces into 3-D solid volumes if the extra lines have the potential to distract or group with other specifiers.

The specifier in relation to the framework

When absolute values are to be communicated:

1. The specifier lines must be no thicker than the level of precision of the tick marks on the axes. In addition, include an inner framework, consisting of a grid pattern, as is discussed below.

2. Keep specifier elements within the framework. If you must have them extend beyond (perhaps to emphasize a point), remember that actual quantitative information will be difficult to extract.

3. If the x axis is more than twice as long as the y axis, include a second y axis on the right of the framework.

Labeling the specifiers

1. Try to avoid using a key or legend. It is better to have the labels directly associated with the specifier elements. Ideally, the label should be closer to the appropriate specifier element than to anything else, allowing the Gestalt principle of proximity to provide grouping. If this is not possible, try having the label at the end of the line, in the wedge, in a bar or in a surface. If this cannot be done, connect the label with the relevant specifier element with an arrow. A key should be used only when a) there are too many specifier elements in too cramped a space or b) the same elements occur in more than two subgraphs in a multiple framework display. Even then, redundant direct labelling is helpful. If a key is used, put it at the top right, within the outer framework, of only a single framework is used, or prominantly above and in the middle of the display if multiple frameworks are used.

2. All labels must be legible, even at reduced sizes; as before, we urge avoiding italics and roman numerals.

3. There should be no more than four labels in a key. If there are, use multiple frameworks (subject to the caveats mentioned above).

4. Use the same font size and style for each member of a set of specifiers comprising a second independent variable (i.e., the parameter) as was used to label the axes(this indicates that the parameter has the same logical status as the axes).

5. If labels are used in a key, make sure the connection between the label and the appropriate specifier element is clear and unambiguous. Associate the label with a segment of the specifier by putting the label closest to the appropriate segment of a superior pattern, and be sure to use a segment long enough so that the pattern is easily identified.

6. Use words that are consistent with those used in the text or commonly used to discuss the topic.

Drawing the Inner Framework

After the outer framework, specifier, and labels have been placed, you are now in a position to draw an inner framework. An inner framework is useful when absolute values are to be read from a graph, given that they allow one to link portions of the specifier to the appropriate labelled pips on the axes.

1. The inner framework should not group with the specifier elements or the labels. This can be ensured by always drawing the inner framework with thinner, lighter lines than those used to draw the other graphic constituents.

2. Make the grain of the inner framework appropriate for the level of precision necessary. A coarse grid will not be of much help if detailed measurements are needed, and a fine grid will only get in the way if only general measurements are needed.

3. Every fifth line of the inner framework (if a grid is used) should be slightly heavier, which will help the reader to track along any single line.

4. The ends of the lines of the inner framework should intersect the outer framework at one and only one place, and this place should be easily seen. This will ensure that the inner framework hooks up clearly to the outer framework, so that it maps specific labelled points on one part of the outer framework to specific points (preferably points that are perceptually isolable) on the specifier elements.

Drawing the Background

First, we recommend avoiding backgrounds if there is the slightest chance that they will impair comprehension. If you insist on drawing one, draw it last, because you want to make sure that it does not interfere with the information-conveying parts of the display. If the background is sufficiently dim or sketchy, it can be drawn first, but you then run some risk of having to re-draw parts of it.

1. Make sure that the background is not too visually dominant. It should be visibly less salient than any other part of the display.

2. Make sure that the background does not draw attention from the display because of its complexity or because parts of it seem to group with parts of the display. If the background is sufficiently dim relative to the display, or of a different degree of fineness of detail, this problem can be avoided. Every element in the background should obviously belong to the background and not to the display.

3. If background figures are used, they should convey a message consistent with the point of the display.

CHAPTER 8: EXTENSIONS AND GENERALIZATIONS

We have so far concentrated on how one analyzes and constructs charts and graphs. But at the beginning of this book we claimed that this focus was largely for methodological reasons, and that the results of our enterprise would in the end have considerably more applicability. That is, we claimed that charts and graphs had the virtue of being highly constrained, and yet of having a wide variety of different possible types. Thus, we expected that we would be led to develop a set of principles and techniques that could be generalized to other kinds of visual displays simply by modifying some of the requirements for charts and graphs proper. Let us first consider how we would extend our approach to the other display types noted in Chapter 2, and then consider a much broader extension of the current project.

Generalizing to other types of displays

The key to generalizing to other types of visual displays is to realize that the system we have developed does not hinge on the precise nature of the graphic constituents. We hoped to illustrate this by using both charts and graphs, in which the frameworks and specifiers have very different forms. The entire system requires only that there be a way of dividing a display into parts that specify different kinds of entities. Once this is done, one can proceed to describe these entities and the relations among them at the level of syntax, semantics, and pragmatics. All of the syntactic principles are applicable to any visual display. That is, the designer of any kind of display must ensure that the marks will be discriminable, must take into account possible distortions introduced by the visual system, must be aware of how marks are grouped together by the visual system, and must take into account the effects of processing priorities and limitations. The same is true for the formal mapping principles. In any display, one must ensure that every mark has only one interpretation and that all necessary marks are present, and one must ensure that the inter-relations among the marks themselves are clear.

In contrast, the semantic and pragmatic principles developed thus far are not directly applicable to all kinds of displays. This does not present severe problems for the semantic principles, however: if a semantic principle discussed thus far is relevant at all, it is applicable as it stands. If a principle is not relevant, it simply should be ignored. Some principles will not be relevant when displays do not involve symbolic representation (i.e., where lines represent via a convention, not via depiction). For example, in many diagrams the display represents solely by depicting an object or part. In this case, the only relevant principles are those that pertain to depictions proper (i.e., representativeness, concept availability). The same is true for the Pragmatic principle of "contextual compatibility". If a display is embedded in a context, it must use terminology consistent with that used in the context and it should neither "tell more or less" than is required in that context. However, the situation is not so simple with the other pragmatic principle, "invited inference". The general idea, that stimulus properties may invite an inference, is applicable for all kinds of visual displays. But precisely how one accomplishes this will vary depending on the kind of display. Not only are some of the principles we describe in Chapter 4 not relevant for many displays (e.g., those not containing axes), but other principles which we have not developed will be relevant. For example, with maps, use of different numbers of elevation rings can convey the impression that a hill is in fact steeper or shallower than it is. We have not begun to work out the principles of invited inference that are relevant for each type of display, but are confident that this can be done.

Let us now consider in more detail how to extend the analytic and generative scheme developed for charts and graphs to other types of displays. In so doing we must delineate the basic-level graphic constituents for each type, and we must note which principles are apt to be irrelevant, and we must propose new principles that might be relevant to specific special cases.

Maps: Let us first consider simple "pure" maps, which contain a depiction of a territory with associated labels. These maps can be composed of a single unit (e.g., a map of a state) or can be composed of a number of sub-maps (e.g., a map of the U.S. showing state boundaries). If the map is divided into subunits, these units are the "basic level constituents" of the analysis. If line widths vary such that relatively small units are nested within areas demarcated by heavier lines, then the largest unit with heavy lines (which is not the entire area) are to be treated as the basic-level unit. (Recall that the basic-level is that which is as general as possible while still having constituent members that are as similar as possible). The relations among the constituents are simple contiguity: regions that abut are organized as representing territories that abut in just that way.

Pure maps of the sort considered above are a rarity, occuring only in special contexts (e.g., globes of the world). Most maps are designed with the intent of conveying specific information about a territory. Road maps tell one about highways, rainfall maps about rainfall, census maps about population, and so on. These maps use conventional symbols as specifiers, relating regions in specific ways. A line is taken to be a road (with a wide yellow one as a superhighway, a narrow red one as a backroad, etc.), a region of dark blue to be one in which rain falls over 300 in. a year, a region of white to be one in which less than 4 in. of rain falls per year, etc. These maps use the depictive component--the territory--as a framework and the lines, regions and so on

301

as specifiers conveying information about specific relations among features of parts of the territory.

Now let us consider slightly more complex maps, in which a visual table is superimposed over the map. In this case, there might be a spike over each location, with taller spikes representing greater populations. Or a circle might be drawn, with its area representing the average yearly rainfall at that location. In this case, the map is serving to label the elements of the visual table (by providing the location which is relevant for that information), and the magnitude of the spike, circle or whatever is interpreted visually as indicating the relative amount of whatever variable is represented.

In the rare cases in which a map represents solely by depiction, the semantic principles concerned with proper pairing of a symbol and a concept are irrelevant. When color, texture, or some other visual property is added to the map to convey information symbolically, either as a specifier or a label, then all of the semantic principles developed previously are relevant. Similarly, when a visual table is imposed over a map, now all of the semantic principles are relevant. In this case, the formal principles are relevant not only to the map, but to its relation to the table.

The pragmatic principles one might want to develop for maps would depend on the specific kind of map being considered. A topological map can be modified to emphasize or de-emphasize height differences (e.g., by spacing of rings); a road map could emphasize or de-emphasize congestion (by varing the size of the marks used to represent roads); a map of population density could emphasize or de-emphasize the unevenness of distribution (by varing the size of the region in which population was averaged over), and so on. It is impossible to work out all such principles beforehand, but the over-riding idea is the same as for charts and graphs: be aware of the distorting effects of the way we describe appearances (as big, little, etc.), and strive to avoid them when they have the potential to mislead.

Diagrams: Diagrams are schematic pictures of objects or entities. A diagram of a machine, for example, represents purely by depiction. A diagram of wind patterns, in contrast, represents symbolically. Many diagrams include both a depictive and symbolic component, as occurs in "exploded view" diagrams in which parts of an object are separated and connected by arrows (the pictures of parts are depictions, the arrows are symbols). The components of diagrams are determined in two ways: first, the actual components of the represented object in part determine how one should analyze the diagram. A portion of the diagram corresponding to a distinct component of the machine will be analyzed as a constituent unit. Second, heavy lines or other perceptual factors (e.g., color differences) may also serve to define a part as a separate unit; even in this case, however, the unit so defined often will correspond to an actual part of the object itself. The relations among constituents will again be ones of spatial contiguity and of functional contingency (how one part can affect another). In many cases, the diagram will not have a distinct specifier; it will merely depict the entities of interest. In others, however, one part may be of particular interest vis-a-vis how it functionally relates ("pairs", to use the term introduced in Chapter 2) two other parts (serving as parts of the framework). For example, a diagram may be intended to show how a given kind of crankshaft fits in an engine. Now, the crankshaft serves as a specifier, and what is important is how it relates to the other components of the engine. Or, in the case of the exploded diagram, the parts are related together by arrows, which serve as specifiers.

Diagrams behave almost exactly like maps in how they represent information, except that they depict some object or entity rather than a territory. Thus, parts can be labeled by words or visual properties, and a visual table can be superimposed over a diagram (e.g., using different colors to show the temperature at each point of an engine). Thus, the comments offered about the applicability of the semantic and pragmatic principles are equally appropriate here.

Visual tables: A visual table is like a numerical table, except that values are represented by visual properties of symbols or depictions. For example, increased amounts of oil could be represented by larger pictures of oil barrels, by bigger blotches, or by darker swatches of gray. The constituents here are the specifiers and labels; if there is a framework, it serves merely as a way of labeling the specifiers. In contrast to graphs, the meaning of the specifiers does not derive from mapping parts of the framework to other parts of the framework. Thus, an analysis of a visual table involves isolating the individual specifiers, and ensuring that they are properly identified (either via recognizing a depiction or associating a label) and that they are properly interpreted (e.g., with bigger shapes representing more of some quantity). In these cases, the horizontal formal mapping principle may not be applicable if labels are directly associated with each specifier; if labels must be extracted via a key or via a framework, then this principle is applicable. Other than this, all of the syntactic, formal and semantic principles described in this book are applicable to these displays. Again, however, the pragmatic principles are less clearly related. To the extent that simple size represents quantity, however, then all of the principles of invited inference developed in Chapter 4 that affect apparent size (e.g., varying irrelevant integral dimensions) will be applicable here.

As should be clear from even this brief treatment, the core of the system we have developed in this book is easily generalizable to other forms of visual displays. All of the display types considered above are less constrained than a high-precision graph, in which points along a specifier must relate together a specific pair of points, one lying on each axis. The principles we needed to consider when constructing such displays encompass those we need to construct good maps, diagrams and visual tables. The principles that dictate how to emphasize a particular point, however, depend in large part on the point itself and the way a given display works.

II. Species of Visual Displays

The project described in this book is an example of how a body of facts, concepts, and theories developed in "pure" research can be brought to bear in the service of dealing with an "applied" problem. One of the reasons it is interesting to engage in this kind of exercise, in developing a technology from a science, is that in the course of developing the technology one often ends up inspiring new "pure" science. This is true in the physical sciences, and it should not be surprising that it is true here. Thus, in this last section of the book we would like to show how this project on charts and graphs feeds into a more general domain, the study of visual representation as a whole.

Let us begin by considering different types of visual displays, using a more general taxonomy than the one just considered. In this taxonomy, we will consider three broad types of uses to which displays can be put. Further, we will use a more general taxonomy than one dividing displays into charts, graphs, maps, diagrams and visual tables; as should have been clear from the foregoing discussion, some of these display types are almost variants on a common form. Consider the taxonomy presented in Table 8.1. The columns of the table correspond to different types of displays. The first type are "intrinsic configurations", where the lines do not refer to anything else. A diagram used in geometry is of this type, as is a purely decorative pattern. The second type are "models", where the lines refer to something else, serving to portray that which is referred to. A drawing of an object, scene, layout, or a map is of this type. The third type are "symbolic" representations, where the lines refer to something else, but that something else is not actually shown (usually because it is an abstract idea or state). Charts, graphs, and abstract "notations" (e.g., Venn diagrams) are of this type. Intermediate cases are of course possible, but these are formed when elements of a display fall into different categories. One example of this is an "exploded diagram", in which pieces of a device are drawn separately with arrows indicating how the pieces fit together; the pieces are models, the arrows are symbols.

INSERT TABLE 8.1 HERE

Now, let the rows of the table correspond to different uses of displays. The top row contains displays that are used merely to illustrate or present information. A drawing used to illustrate a rhombus, to show the layout of a house, or to indicate rising prices by a rising line in a bracket (i.e., a graph) fall into the three columns, being examples of intrinsic configurations, models, and symbolizations, respectively. The next row corresponds to displays that are used to help one solve a problem. In this case one does more than simply extract information from a display; one uses the display to help one reason through to a solution to the problem represented in the display. A diagram used to prove a theorem in geometry, a picture of pulleys used to anticipate what will happen when one pulls the rope a certain distance, and Venn diagrams used to solve logic problems fall into the three columns along this row. Finally, the last row contains displays that are generated when one is trying to discover the best way to formulate a problem in the first place. In this case one often may generate numerous different displays, considering the implications of each, before making one that seems to provide insight into how to look at a problem. Presumably some of the diagrams Euclid drew belong in the first column, some of the images Einstein reported definitely belong in the second column (e.g., of himself riding on a beam of light, when he first began to ponder relativity), and scribblings created by untold numbers of mathematicians belong in the third column in this row.

The foregoing taxonomy is interesting in part because it defines a hierarchy of principles for visual display design. In the first row we have the syntactic principles that dictate how marks will be organized, encoded, and retained in active memory. For example, marks near each other will tend to be grouped together, marks that are drawn with heavier lines will tend to be noticed sooner, and too many perceptual units will be difficult to apprehend (see Chapter 3). These sorts of principles apply to all of the cells in this row. In addition, in the second two cells we have semantic principles that dictate how patterns of marks will be interpreted as conveying meaning (intrinsic configurations do not refer to anything else, and thus are not interpreted semantically). The semantic principles that are appropriate for models are straightforward; they deal with the way pictures are seen as resembling objects. The external and internal mapping principles are applicable here, as are the principles of representativeness and concept availability. The rightmost cell in the first row inherits not only the syntactic principles that are relevant to the first cell, and the semantic princples that are relevant to the second, but also adds yet another layer of semantic principles to these principles. The semantic principles that apply only to symbolic displays are more complex, focusing on how variations in marks (e.g., size, color, texture) map into conceptual dimensions (see Chapter 4). For example, bigger marks will be interpreted as representing "more" of some thing.

INSERT FIGURES 8.1 AND 8.2 HERE

The principles assigned to each cell in the first row are inherited by the corresponding cells in both rows beneath them. That is, these principles are equally valid for diagrams used merely to convey information, used to solve a problem, or used to help formulate a problem. In addition, in the second row we add another set of principles. These displays are not simply read, but are actively processed in the course of using them to solve problems. Thus, we can add a second set of principles here, which specify how the various displays can be processed and the best ways of processing them. Imagery would appear to be a key means by which these sorts of displays are used. Moving across the columns: Imagery is often reported when subjects try to solve geometry problems, such as proving that two regions of a diagram have the same area. In this case, parts may be imaged and rotated, shifted across the page, expanded or reduced in size, and compared to other parts. For example, consider Figure 8.1. Does the inner square have half the area of the outer one? One way to solve this problem is to fold the corners of the outer square so that the tips meet in the center, and to "see" in the image that in so doing, one neatly just covers the inner square. Kosslyn (1980) specifies the principles that constrain how such images operations can proceed. The principles of imagery processing that are relevant here generalize to the other two cells. Moving to the middle cell, imagery is often used when one anticipates how a model would look when in motion. In this case, imagery is used to conduct a kind of "simulation" on the diagram, with the aim of mimicking the corresponding actual event (such as by imaging how gears will interact when the first one in a series is twisted clockwise). For example, consider Figure 8.2. If the leftmost gear is twisted clockwise, which way will the rightmost gear move? Finally, in the last cell in this row, imagery is used to manipulate symbols. For example, Venn diagrams are sometimes reported to be "seen" to slip in and out of one another and swinging about in various ways when one is trying to discover if a certain conclusion follows from a set of premises. In all three cells of this row, then, we not only have the principles inherited from the first row, but we have additional principles that dictate how displays should best be processed to achieve certain ends. It seems safe to say that we have just begun to make progress in discovering and formulating these principles.

One of the consequences of considering together the two sets of principles, those derived from studying charts and graphs and those derived from the study of mental imagery, is that they will interact. That is, in designing a diagram to be used as an aid to solving problems one must consider not only factors that pertain to the diagram itself, but also factors pertaining to how easily imagery can be used in conjunction with the diagram. For example, by varying the line weights some parts of a diagram could be emphasized over others, or some organizations of the figure made more salient than others. Depending on what the illustrator wants the viewer to do with a diagram when using it as an aid to solving problems, different stimulus factors can be varied to encourage different imagery manipulations. For example, in Figure 8.1, if the small triangles that are to be imaged folding are emphasized with heavier lines, this might encourage people to attempt to image those parts being manipulated in different ways. This conjecture could be studied directly, and in fact the entire realm of diagram design for problem solving is ripe for study.

Finally, the bottom row of the table inherits all of the principles that pertain to the previous two uses of displays. But now we must add another set of principles that are specific to these kinds of displays, namely those that pertain to how displays should be created to help formulate a problem. These principles will be intimately tied up with principles of creative thinking in general, and remain a mystery at the present writing.

Thus, it is clear that the attempt to use the available data, concepts and theories in an applied setting has not been of mere technological interest. The principles and theories we have developed can easily be used as building blocks in more general projects addressing broader issues of representation and of use of visual information.

III. Conclusions

In this book we have attempted to accomplish three things: First, specifically we have tried to discover what makes a good (or bad) chart or graph. This goal has resulted in an analytic scheme which one can use to diagnose the problems with a given chart or graph and a set of guidelines to help one construct good charts and graphs in the first place. Second, we have tried to develop a general conception of what is going on in the head of a reader when he or she is extracting information from charts and graphs. This theory was useful in part in its role of providing heuristic guidelines for the construction of good charts and graphs. Finally, we have tried to show how this particular project is just the tip of an iceberg; not only can our specific guidelines be generalized to other kinds of visual displays, but the theoretical framework we have developed can serve as the foundation for scientific work in more general problems of visual representation. We are painfully aware of the deficiencies of our accomplishments on all three counts, but are encouraged by how easily our accomplishments were achieved and how clearly the issues and questions have presented themselves. We hope that this book provides both practical tools for illustrators, and inspiration to other researchers to continue to demonstrate that scientific approaches to psychology have much to offer society at large.

REFERENCES

Abramov, I. and Gordon, J.: Vision. Carterette, E. and Friedman, M., Eds., *Handbook of Perception*, Vol. 3, 1974, New York: Academic Press.

Aidley, D.J: *The Physiology of Excitable Cells*, 1971, Cambridge: University Press.

Baird, J.C.: *Psychophysical Analysis of Visual Space*, 1971, Cambridge: University Press.

Baird, J.C. and Noma, E.: *Fundamentals of Scaling and Psychophysics*, 1978, New York: John Wiley & Sons.

Bertin, J.: *Semiologic Graphique: Les Diagrammes - Les Reseaux - Les Cartes*, 1967, The Hague: Mouton.

Biderman, A.D.: *Kinostatistics for social indicators*. Educ. Broadcast. Rev. 5:13-19.

Campbell, F.W. and Robson, J.G.: Application of Fourier Analysis to the Visibility of Gratings. *Journal of Physiology*, 1968, 197, 551-556.

Clark, H.H., and Card, S.K.: The role of semantics in remembering complex sentences. *Journal of Experimental Psychology*, 1969, *82*, 545-552.

Clark, H.H., and Clark, E.V.: Semantic distinctions and memory for complex sentences. *Quarterly Journal of Experimental Psychology*, 1968, *20*, 129-138.

Cleveland, W.S., Diaconis, P., and McGill, R.: Variables on Scatterplots Look More Highly Ordered when the Scales are Increased. Unpublished manuscript, 1981.

Cobb, P.W. and Moss, F.K.: The Four Variables in the Visual Threshold. *Franklin Institute Journal*, 1928, 205-831.

Conover, D.W.: The Amount of Information in the Absolute Judgement of Munsell Hues. WADC-TN58-262, 1959, Wright Air Development Center, WPAFB, Ohio.

Cornsweet, T.N.: The Staircase-Method in Psychophysics. *American Journal of Psychology*, 1962, 75, 485-491.

De Valois, R.L. and Marrocco, R.T.: Single Cell Analysis of Saturation Discrimination in the Macaque. *Vision Research*, 1973, 13, 701-711.

Duncan, J. and Konz, S.: Legibility of LED and Liquid-Crystal Displays. *Proceedings of the S.I.D.*, 1976, 17(4), 180-186.

Engen, T.: Psychophysics I, Discrimination and Detection, *Woodworth and Schlosberg's Experimental Psychology, Vol. 1: Sensation and Perception*, J.W. Kling and L.A. Riggs, Eds., 1972, New York: Holt, Rinehart and Winston.

Field, J. and Magoun, H.W. (Eds.): Handbook of Physiology, Section 1: Neurophysiology, Vol. 1, 1959, American Physiological Society, Washington, D.C.

Fletcher, D.: Matching Operator's Eyes with Machine Displays. Digital Design, 1972, 2(11),, 42-43.

Goldhamer, H.: The Influence of Area, Position, and Brightness in the Visual Perception of a Reversible Configuration. American Journal of Psychology, 1934, 46, 189-206.

Graham, C.: Area, Color and Brightness Difference in a Reversible Configuration. Journal of General Psychology, 1929, 2, 470-481.

Guilford, J.P.: A Generalized Psychophysical Law. Psychological Review, 1932, 39, 73-85.

Harris, J.D.: The Decline of Pitch Discimination with Time. Journal of Experimental Psychology, 1952, 43, 96-99.

Hartley, J.: Designing Instructional Text. London: Kogan Page, 1978.

Helmhotlz, H. von: Handbuch der Physiologischen Optik, Hamburg and Leopzig, Voxx, 1866.

Hochberg, J.: Perception: Color and Shape in Experimental Psychology (Lking and Riggs, Eds.), Holt, Rinehart and Winston Inc., New York, 1971.

Hochberg, J.: Organization and the Gestalt Tradition in Handbook of Perception, Vol. 1 (Carterrette and Friedman, Eds.), Academic Press, New York, 1974.

Indow, T. and Stevens, S.S.: Scaling of Saturation and Hue, Perception and Psychophysics, 1966, 1, 253-271.

Jones, L.A. and Lowny, E.M.: Retinal Sensibility to Saturation Differences. Journal of the Optical Society of America, 1926, 13, 25-37.

Kellogg, W.N.: An Experimental Comparison of Phychophysical Method. Archives of Psychology, 1929, Vol. 17, No. 106, 1-86.

Kelly J.J. and Bliss, W.D.: A Psychophysical Evaluation of the Accuracy of Shape Discrimination as an Aircraft Landing Aid. Human Factors, 1971, 13(2), 191-193.

Kincha, R. and Smyzer, F.: A Diffusion Model of Perceptual Memory. Perception and Psychophysics, 1967, 2, 219-229.

Kling, J.W. and Riggs, L.A. (Eds.): Experimental Psychology, 1971, New York: Holt, Rinehart and Winston.

Kohler, W. and Adams, P.A.: Perception and Attention in Documents of Gestalt Psychology, (Henle, Ed.), University of California Press, 1961.

Kunnapas, T.: Experiments on Figural Dominance, Journal of Experimental Psychology, Vol. 53, No. 1, 1957.

Land, E.H.: Experiments in Color Vision. Scientific American, 1959.

Lowry, . The Photometric Sensibility of the Eye and the Precision of Photometric Observations. Journal of the Optical Society of America, 1931, 21, 132-136.

Luce, R.D., Bush, R.R. and Glanter, E.: Handbook of Mathematical Psychology, 1963, New York: John Wiley and Sons.

Macdonald-Ross, M.: 1978. Research in graphic communications. IET Monog. 7. Reprint of "Graphics in texts", in Rev. Res. Educ. Vol. 5, 1977.

Marriott, F.H.C.: The Interpretation of Multiple Observations, 1974, New York: Academic Press.

Mil-STD-1472B: Human Engineering Design Criteria for Military Systems Equipment and Facilities. Washington, D.C.: U.S. Department of Defense, 31 December, 1974.

Miller, G.A.: Sensitivity to Changes in the Intensity of White Noise and its Relation to Loudness and Masking. Journal of the Acoustical Society of America, 1947, 19, 609-619.

Moon, P., and Spencer, D.D.: Visual Data Applied to Lighting Design. Journal of the Optical Society of America, 1944, 34, 605.

Murrell, R.H.: Ergonomics, 1965, London: Chapman and Hill.

Ono, H.: Difference Threshold for Stimulus Length Under Simultaneous and Nonsimultaneous Viewing Conditions. Perception and Psychophysics, 1967, 2, 201-207.

Orbison, W.D.: Shape as a Function of the Vector Field. American Journal of Psychology, 1939, 52, 31-45.

Oyama, T.: Figure-ground Dominance as a Function of Sector Angle, Brightness, Hue, and Orientation. Journal of Experimental Psychology, 1960, Vol. 60, No,. 5, 299-305.

Panek, D.W. and Steven, S.S.: Saturation of Red: A Prothetic Continuum. Perception and Psychophysics, 1966, 1, 59-66.

Poulton, E.C.: Searching for newspaper headlines printed in capitals or lower case letters. Journal of Applied Psychology, 1967, 51, 417-425.

Priest, I.G. and Brickwedde, F.G.: The Minimum Perceptible Colormetric Purity as a Function of Dominant Wavelength. Journal of the Optical Society of America, 1938, 28, 133-139.

Rouse, W.B.: The Effect of Display Format on Human Perception of Statistics, Proceedings of the Tenth Annual Conference on Manual Control, Wright Patterson AFB, April 1974.

Siegel, M.H.: Discrimination of Color IV. Sensitivity as a Function of Spectral Wavelength, 410 Through 500 mμ. Journal of the Optical Society of America, 1964, 54, No. 6, 821-823.

Sloan, L.L.: Measurement of Visual Acuity. Archives of Opthalmology, 1951, 45, 704-725.

Smith, S.L.: Letter Size and Legibility. Human Factors, 1979, 21(b), 661-670.

Spencer, H.: The Visible Word. London: Lund Humphries, 1969.

Stewart, T.F.M.: Displays and the software interface. Applied Ergonomics, 1976, 7, 137-146.

Talbot, S.A. and Gessner, U.: Systems Physiology, 1973, New York: John Wiley and Sons.

Taves, E.H.: Two Mechanisms for the Perception of Visual Numerosisness. Archives of Psychology, 1941, 37, No. 205.

Thomas, J.P.: Spatial Resolution and Spatial Interaction. Caterette and Priedman, Eds., Handbook of Perception, Vol. 5, 1975, New York: Academic Press.

Tinker, M.A. and Paterson, D.G.: Studies of typographical factors influencing speed of reading: VII. Variation in colour of print and background. Journal of Applied Psychology, 1931, 15, 471-479.

Titchener, E.B.: An Outline of Psychology, New York, Macmillan, 1902.

Tufte, E.R.: 1977. Improving data display. Dept. Stat., Univ. Chicago.

Tufte, E.R.: 1978. Data graphics. First Gen. Conf. Soc. Graphics, Leesburg, VA.

Tukey, J.W.: 1977. Exploratory Data Analysis. Reading, Mass: Addison-Wesley.

Veniar, F.A.: Difference Thresholds for Shape Distortion of Geometrical Squares. The Journal of Psychology, 1948, 26, 461-476.

Wainer, H.: 1977. Data Display--Graphical and Tabular. Hackensack, NJ: NCCD.

Wainer, H.: 1978. Graphical display. Presented at Eur. Math. Psychol. Assoc., Uppsala, Sweden.

Wainer, H.: Reply: The American Statistician, 1981, 35, 57-58.

Wainer, H., Francolini, C.: 1980. An empirical inquiry into human understanding of two variable color maps. Am Stat. 34:81-93.

Wainer, H., Thissen, D.: 1979. On the robustness of a class of naive estimators. Appl. Psychol. Meas. 4:543-51.

Wainer, H. and Thissen, D.: Graphica Data Analysis. Ann. Rev. Psychol. 1981, 32:191-241.

Waller, R.H.W.: Notes on Transforming, No. 4. Milton Keynes: Open University, Institute of Educational Technology, 1977. Revised and reprinted in J. Hartley (Ed.). The Psychology of Written Communication. London: Kogan Page, 1980.

Warner, J.D.: *A Fundamental Study of Predictive Displays Systems*, NASA CR-1274, 1969.

Weber, E.H.: Der Tastsinn und das Gemeingefuhl, R. Wagner, Ed., *Handwortenbuck der Physiologie*, 1969, 6, 13-61.

Wertheimer, M.: *Laws of Organization in Perceptual Forms in a Source Book of Gestalt Psychology* (Ellis, W.D., Ed.), Routledge & Kegan Paul Ltd., London 1938.

Wever, E.G.: Attention and Clearness in the Perception of Figure and Ground. *American Journal of Psychology*, 1928, 40, 51-74.

Whalley, P.C., and Flemming, R.W.: An experiment with a simple recorder of reading behavior. *Programmed Learning and Educational Technology*, 1975, *12*, 120-123.

Wickelgren, W.A.: Associative Strength Theory of Recognition Memory for Pitch. *Journal of Mathematical Psychology*, 1969, 6, 14-61.

Wickens, C.D., and Kessel, C.: The Effects of Participatory Mode and Task 19 Workload on the Detection of Dynamic System Failures. Proceedings of 13th Annual Conference on Manual Control, 1977, Cambridge, MA.

Wier, C.C., Jesteadt, W., and Breen, D.M.: A Comparison of Methods of Adjustment and Forced-Choice Procedures in Frequency Discrimination. *Perception and Psychophysics*, 1976, 19, 75-79.

Woodworth, R.S.: *Experimental Psychology*, 1938, New York: Holt.

Wright, P.: New York: Cambridge University Press, in press.

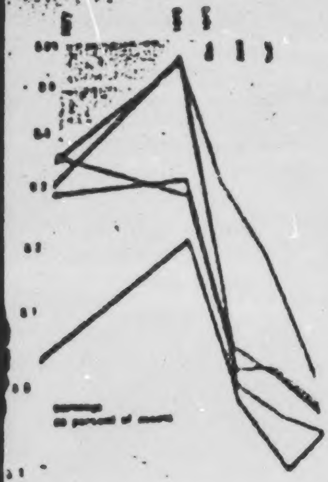
Wright, P. and Threlfall, S.: Readers' expectations about format influence the usability of an index. *Journal of Research Communication Studies*, 1980, *2*, 99-106.

Wundt, W.: *Outline of Psychology* (4th German edition, translation by C.H. Judd), Leipzig Englemann, 1902.

Wyszecki, G. and Stiles, W.S.: *Color Science*, 1967, New York: John Wiley and Sons.

Zanforlin, M.: Some Observations on Gregories Theory of Perceptual Illusions. *Quart Journal of Experimental Psychology*, 1967, 29, 193-197.

FIVE THAT ARE FALTERING



[illegible]

... are generally higher ... [illegible] But they ... [illegible] regular ... [illegible]

... [illegible] generally ... [illegible]

Fig 1.1

g technique, which parently involve a mponent than it does earched. They may hologically

ults of niche ... that reach ...

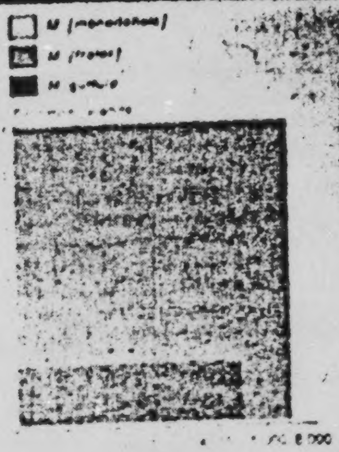


fig 1.2

... a few ... Under ... is far less of ... all out of ... by depleting resource

... interspecific competition is ... often reflected in peaceful ... existence or in permanent exclusion than in running warfare.

ting ... than ... the phrase ... the ...ence." The ... age of fighting is

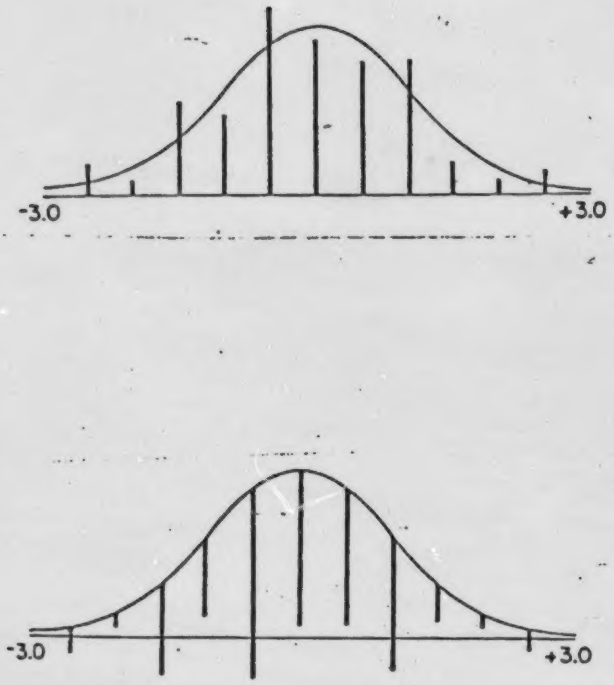
... exclusion.

Consider, then, the likelihood of de-

Why so late?

Most of the detailed evidence for the importance of interspecific competition in nature has been gathered in the last twenty years. Yet no special

(Chapter) 1
Figure 1.3



1.3.
Figure 7. A conventional histogram (top), and
a "hanging histogram" (bottom).

Fig 1,3

2.1

TABLE II-1:

Domain Scale	Range Scale	Example	Examples of Information Available
Nominal	Nominal	States w/ and w/out Capital Punishment	Frequencies in cells; named things sorted into classes on all/none basis.
Nominal	Ordinal	States by rank in coal production	Allow N(N-1)/2 inequalities statement to be made.
Nominal	Interval	Students by score on acheivement test	Map N things into an infinite amount of classes. Comparison of differences.
Nominal	Ratio	States by coal production	Nonarbitrary zero point. Ratio comparison of items possible.
Ordinal	Ordinal	Ranked oil production by ranked coal production	Relative ranks. Comparison of disparties in ranks.
Ordinal	Interval	Rank in classes by acheivement test score	Numerical assignment of the relative position of some characteristic of an item.
Ordinal	Ratio	Ranked coal production by oil production	Relative difference or ratios of oil for different ranks.
Interval	Interval	Math acheivement score by english acheivement score	Differences on both dimensions. Difference in one dimension as a function of difference in other.
Interval	Ratio	Acheivement scores by hours/week of TV watching	Mapping into nonarbitrary zero point scale specified relationships.
Ratio	Ratio	Coal production by oil production	Absolute amounts, differences in amounts, ratios, for both dimensions; value of one dimension as a function of other

Table 2.1

Figure 2.1

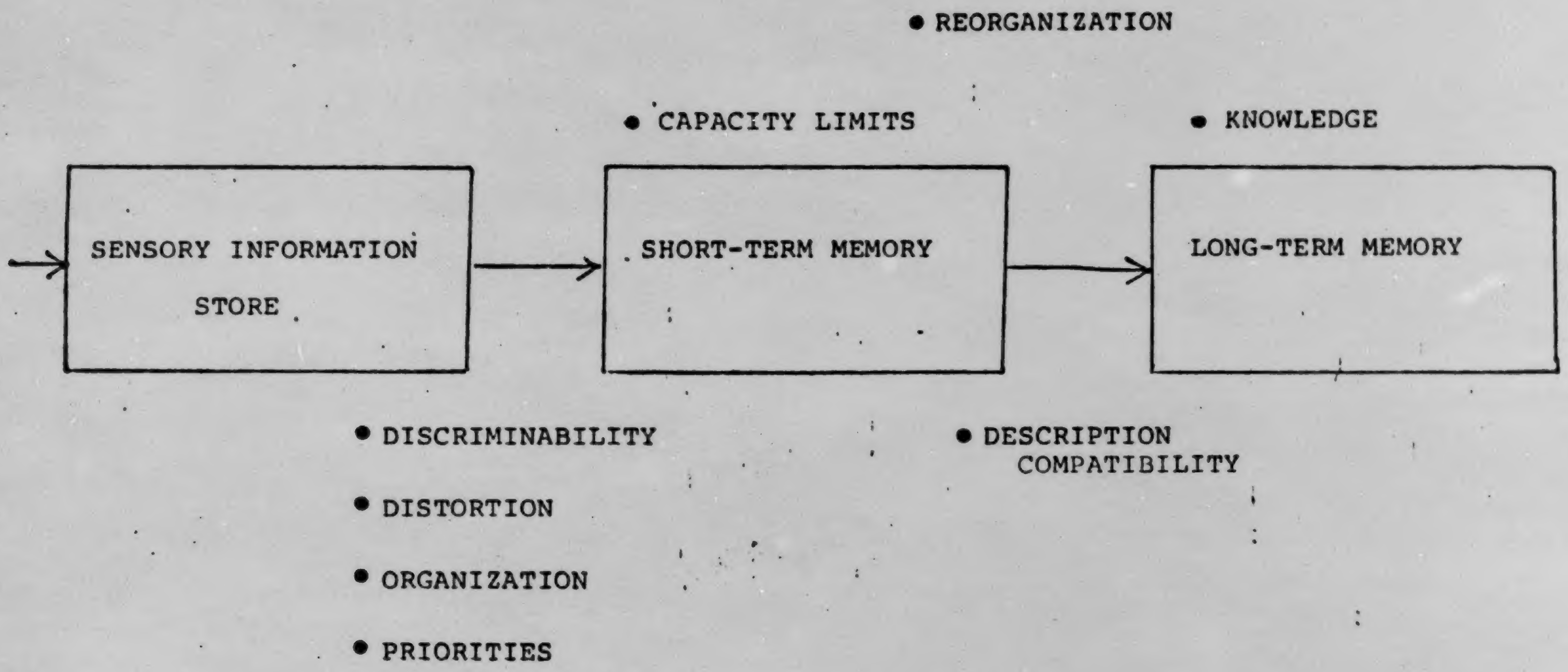
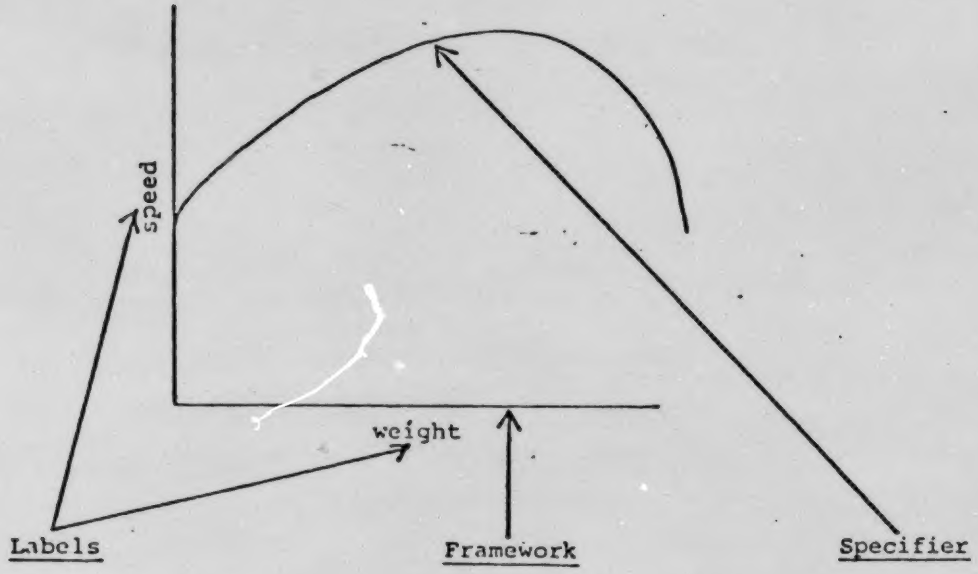
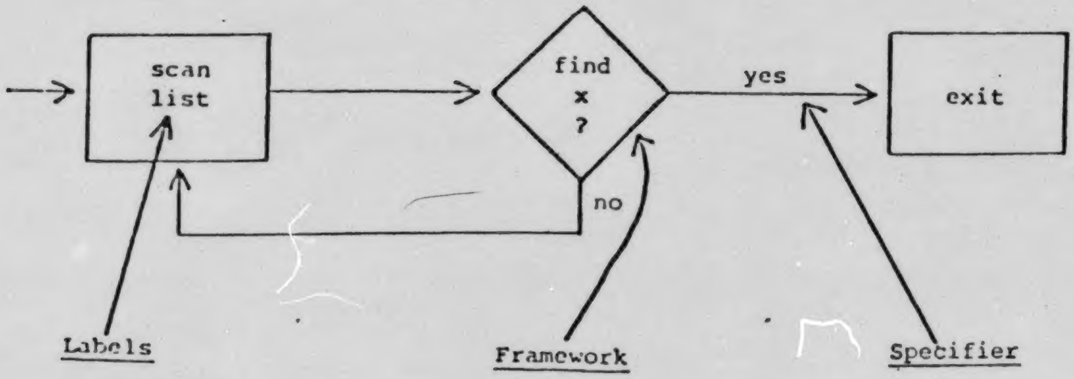

• REORGANIZATION
• CAPACITY LIMITS
• KNOWLEDGE
SENSORY INFORMATION STORE
SHORT-TERM MEMORY
LONG-TERM MEMORY
• DISCRIMINABILITY
• DISTORTION
• ORGANIZATION
• PRIORITIES
• DESCRIPTION COMPATIBILITY

Figure 1: Examples of the Three Basic-level Constituents of a Chart or Graph


speed
weight
Labels
Framework
Specifier

A. GRAPH


scan list
find x ?
yes
no
exit
Labels
Framework
Specifier

B. CHART

Fig. 2.2

~~FIGURE 1~~

AVERAGE STOPPING DISTANCE

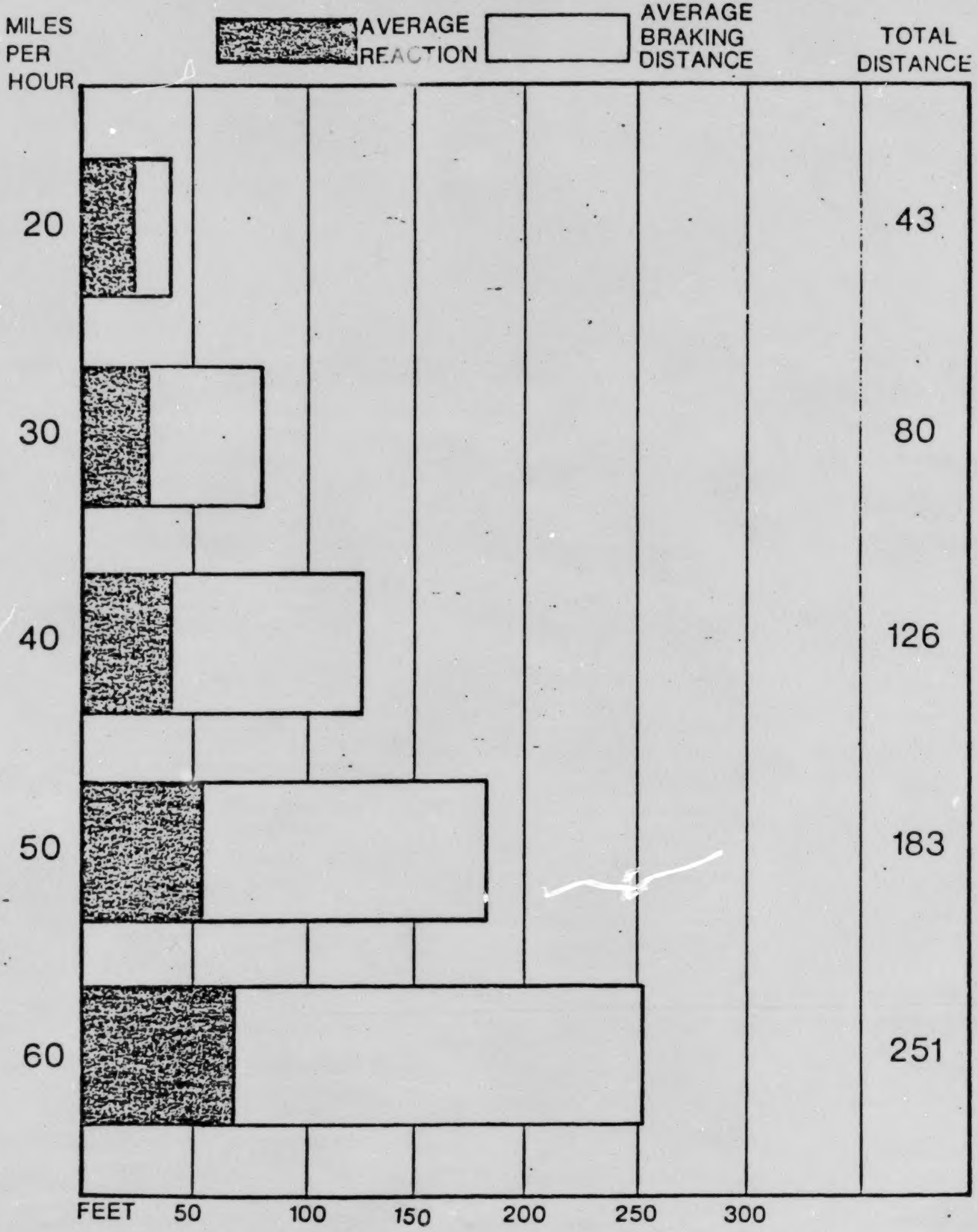
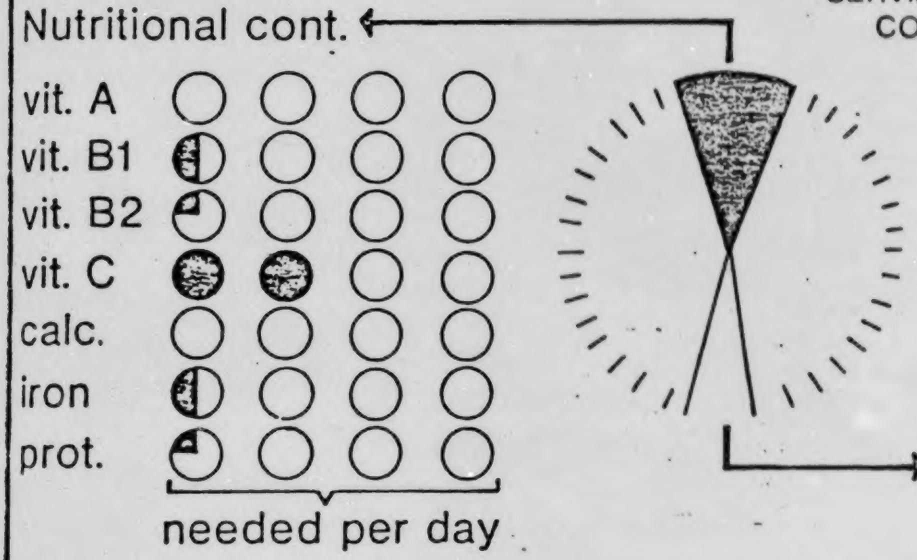


FIG 2.3

~~FIGURE 2~~

NUTRITIONAL INFORMATION PER SERVING

SERVING SIZE	ONE CUP
SERVINGS PER CONTAINER	8



protein	4g
fat	0g
carbohy.	38g

FIG 2.4

Table 1

Summary of Cited Recommendation for size of Displayed Letters

Condition	Visual Angle In Minutes of Arc
Normal Acuity (Snellen E Chart)	5
Reasonable Size (of numerals) (Murrell, 1965, Fletcher, 1972)	10
Preferred Size (of numerals (Duncan and Konz, 1976)	23
MIL-STD-1472B (1974)	
General Labels, good viewing	16+
Noncritical data	6-24
Critical data, fixed position	
High Luminance	12-25
Low Luminance	19-37

Table 3,1

Table 2

CONVERSION FACTORS FOR LUMINANCE UNITS

Units	Foot-lamberts	Lamberts	Milli-lamberts	Candles per square inch	Candles per square foot	Candles per square centimeter
ft.-L........	....	1.076×10^{-3}	1.076	2.21×10^{-3}	3.18×10^{-1}	3.43×10^{-4}
L............	9.29×10^{2}	....	1.0×10^{3}	2.054	2.96×10^{2}	3.18×10^{-1}
mL..........	9.29×10^{-1}	1.0×10^{-3}	....	2.054×10^{-3}	2.957×10^{-1}	3.183×10^{4}
c/in²..........	4.52×10^{2}	4.87×10^{-1}	4.87×10^{2}	....	1.44×10^{2}	1.55×10^{-1}
c/ft²..........	3.14	3.38×10^{-3}	3.38	6.94×10^{-3}	....	1.076×10^{-3}
c/cm²........	2.92×10^{3}	3.14	3.14×10^{3}	6.45	9.29×10^{2}	....

Note: Value in units in left-hand column times conversion factor equals value in units shown at top of column.

327 TABLE 3.2

Table 3,3

ROD AND CONE VISION OF THE HUMAN EYE

	Cone	Rod
Distribution	(ca. 7 million)	(ca. 120 million)
Retinal location	Concentrated at center, fewer in periphery	General in periphery, none in fovea
Neural processing	Discriminative	Summative
Peak wavelength	555 nm	505 nm
Luminance level	Daylight (1 to 10^7 mL)	Night (10^{-6} to 1 mL)
Color vision	Normally trichromatic	Achromatic
Dark adaptation	Rapid (ca. 7 min)	Slow (ca. 40 min)
Spatial resolution	High acuity.	Low acuity
Temporal resolution	Fast reacting	Slower reacting

TABLE 3,3

Table 4 3.4

Differential Sensitivities for Retinal Variables

Retinal Variable	Dimension	Differential Sensitivity (%)	Method	Stimulus Range	Standard	Viewing Conditions	No. of Subjects	Reference
	line length	4.1	limits	3.4-6.6 cm	5 cm			Ono (1967)
				8.4-11.6	10	nonsimultaneous	60	
				13.4-16.6	15			
Size	Area	6.0	-	-	-	-	-	Baird (1969)
	Numerosity	20.4	Constant^{+} Stimuli	9-15 dots	15 dots			
				13-25	25			
				26-50	50	nonsimultaneous	5	Taves (1941
				70-100	100			
				120-180	180			
	Hue	*	Constant Stimuli	410-500 mμ	**	simultaneous	3	Siegel and Dimmick (1962)
				510-630 mμ				Siegel (1964)
Color	Saturation	2.0	Constant Stimuli	-	25% purity^{++}			
					35			
					50	simultaneous	8	Paneck and Stevens (1966)
					65			Indow and Stevens (1966)
					80			
	Brightness	1.4	Adjustment	-	0.62-224.0 ml	simultaneous		Lowry (1941)

(cont.) 3.4

3.4

Table 4 (cont.)

Retinal Variable	Dimension	Differential Sensitivity (%)	Method	Stimulus Range	Standard	Viewing Conditions	No. of Subjects	Reference
Shape	Distortion of Square	1.37	Single[t] Stimuli	25 x 20 cm to 25 x 30 cm	25 x 25 cm	nonsimultaneous	5	Veniar (1948)
	Distortion of Diamonds	4.8	Absolute[tt] Judgments	.925-1.075 (height/width ratio)	height/width ratio of 1.00	nonsimultaneous	20	Kelly and Bliss (1971)

+ JND calculation differed from normal convention

+ For the Paneck and Stevens experiment

* See Figure 11

* Standard varied in 10 mμ steps along total stimulus range

t For a description of this method, see Woodworth (1938)

t Similar to constant stimuli

TM 3.4

3.5

Table 5

NINE EQUALLY-DISCRIMINABLE SURFACE COLORS

Hues	Code number	Munsell book number	Excitation purity	Dominant wavelength
1	1.5	3R	37.2	629
2	3	9R	65.8	596
3	5.5	9YR	81.8	582
4	8.5	1GY	76.0	571
5	11.5	3G	27.5	538
6	15	7BG	35.0	491
7	18	9B	56.5	481
8	20.5	9PB	52.7	460
9	24	3RP	36.5	510

333 [illegible] 3.5

TABLE ~~6~~ 3.6

POWER LAW EXPONENTS FOR LINE LENGTH

No.	Exponent	Variability* Measure	Method	Stimulus Range (cm)	Standard (cm)	Location of Std in Range	Modulus	No. of Subjects	Reference
1.	1.07	-	magnitude estimation	2.1-15.9	-	-	-	-	Bjorkman, Stranger (1960)
2.	.78	-	magnitude estimation	.3-17.7	-	-	-	-	Bjorkman, Stranger (1960)
3.	1.11	-	ratio estimation^T	1.3-2.75	N.A.**	N.A.	N.A.	10	Ekman, Junge (1961)
4.	1.00	-	magnitude production	1.3-254.0	13.5	Low	10	10	Stevens, Guirao (1963)
5.	.98	Oxy = .05	magnitude estimation (based on apparent length)	1.3-83.8	8.9	Low	10	10	Teghtsoonian (1965)
6.	1.02	Oxy = .03	magnitude estimation (based on physical length)	1.3-83.8	8.9	Low	10	10	Teghtsoonian (1965)
7.	1.07	O = .18	magnitude estimation	2.4-9.3	7.7	High	10	36	Rule (1966)
8.	.98	$\sigma x^2 y$ = .00010	magnitude estimation (of circle diameters)	2.0-10.2	10.2	High	not assigned	40	Stanley (1967)
9.	.97	$\sigma x^2 y$ = .00008	magnitude estimation (of vertical lines)	2.0-10.2	10.2	High	not assigned	40	Stanley (1967)

T A modification of the method of constant sums

* The different measures of variability used by investigators are: σxy - sample of std dev from regression (log-log plot), σ - std dev for distribution of individual subject exponents, $\sigma x^2 y$ - residual variance about regression (log-log plot), R - range of individual subject exponents, Rw - width of range of individual subject exponents

** N.A. - parameter is Not Applicable

TABLE 3.6

3.6

TABLE 6
(Continued)

POWER LAW EXPONENTS FOR LINE LENGTH

No.	Exponent	Variability* Measure	Method	Stimulus Range (cm)	Standard (cm)	Location of Std in Range	Modulus	No. of Subjects	Reference
10.	1.14^t	Rw = .07	magnitude estimation	20.3-185.4	30.4	Low	12	18	Miller, Sheldon (1969)
11.	1.00^t	Rw = .24	magnitude estimation	20.3-185.4	91.4	Middle	36	18	Miller, Sheldon (1969)
12.	1.04^t	Rw = .35	magnitude estimation	20.3-185.4	152.4	High	60	18	Miller, Sheldon (1969)
13.	.94	R = .73, 1.39	magnitude estimation	1.27-20.32	none	N.A.**	none	24	Duda (1975)

[t] Each stimulus was a group of six parallel horizontal lines; lengths within a stimulus group were uniformly distributed with a range of 30 cm. Subject's estimated average length for group.

* The different measures of variability used by investigators are: σxy - sample std dev from regression (log-log plot), σ - std dev for distribution of individual subject exponents; σx^2y - residual variance about regression (log-log plot), R - range of individual subject exponents, Rw - width of range of individual subject exponents

** N.A. - parameter is Not Applicable

TABLE 3.6

3.7

TABLE 8

POWER LAW EXPONENTS FOR AREA OF VARIOUS FIGURES

No.	Exponent	Variability* Measure	Method	Stimulus Ratio Max Area/Min Area	Standard Std Area/ Min Area	Location of Std in Range	Modulus	No. of Subjects	Reference
Circles									
1.	.86	-	ratio setting[tt]	9.0	N.A.	N.A.	N.A.	5	Ekman (1958)
2.	.96	-	ratio estimation[tt]	7.0	N.A.	N.A.	N.A.	-	Bjorkman, Strange (1960)
3.	1.20	-	ratio estimation[tt]	26.6	N.A.	N.A.	N.A.	-	Bjorkman, Strange (1960)
4.	.78	-	ratio estimation[tt]	49.0	N.A.	N.A.	N.A.	-	Bjorkman, Strange (1960)
5.	.80[t]	-	magnitude estimation	2.30	-	Middle	100	33	Ekman, Lindman, William-Olsson (1961)
6.	.98	-	magnitude estimation	2.1	1.0	Low	1	10	Ekman, Junge (1961)
7.	1.05	-	magnitude estimation	4.5	1.0	Low	1	10	Ekman, Junge (1961)
8.	.99	-	magnitude estimation	9.5	1.0	Low	1	10	Ekman, Junge (1961)

[t] Squares and circles as stimuli (data was pooled)

[tt] Ratio setting is a modification of fractionation. Ratio estimation is a modification of the method of constant sums

* The different measures of variability used by investigators are: σxy - sample std dev from regression (log-log plot), σ - std dev for distribution of individual subject exponents, $\sigma x^2 y$ - residual variance about regression (log-log plot), R - range of individual subject exponents, W - width of range of individual subject exponents

** N.A. - parameter is Not Applicable

TABLE 7
(Continued)

POWER LAW EXPONENTS FOR AREA OF VARIOUS FIGURES

No.	Exponent	Variability* Measure	Method	Stimulus Ratio Max Area/Min Area	Standard Std Area/ Min Area	Location of Std in Range	Modulus	No. of Subjects	Reference
Circles									
9.	1.03	$\sigma yx = .06$	magnitude estimation (physical area)	81	25	Middle	10	10 grad students	Teghtsoonian (1965)
10.	.76	$\sigma yx = .05$	magnitude estimation (apparent size)	81	25	Middle	10	10 grad	Teghtsoonian (1965)
11.	1.03	$\sigma = .23$	magnitude estimation	210	51	Middle	10	36 undergrads	Rule (1966)
12.	.70	-	magnitude estimation	121	-	-	-	-	Manhour, Hosman (1968)
13.	.69	-	magnitude estimation	1000	none	N.A.	N.A.	-	M. & R. Teightsoonian (1971)
14.	.81	-	magnitude estimation	4.7	none	N.A.	N.A.	-	Vogel, Teightsoonian (1972)
15.	.58	$\rho > .99$	magnitude estimation (apparent size)	179	11.0, 17.4 always present	Low	10,100	4	MacMillan, [illegible] et al. 1974 experiment 1
16.	.55	$\rho > .99$	magnitude estimation (apparent size)	3075	48.0, 75.5	Low	10,100	4	

[tt] Ratio setting is a modification of fractionation. Ratio estimation is a modification of the method of constant sums

* The different measures of variability used by investigators are: σxy - sample std dev from regression (log-log plot), σ - std dev for distribution of individual subject exponents, $\sigma x^2 y$ - residual variance about regression (log-log plot), R - range of individual subject exponents, Rw - width of range of individual subject exponents, ρ - correlation between log ϕ & log ψ

TABLE 7
(Continued)

POWER LAW EXPONENTS FOR AREA OF VARIOUS FIGURES

No.	Exponent	Variability* Measure	Method	Stimulus Ratio Max Area/Min Area	Standard Std Area/ Min Area	Location of Std in Range	Modulus	No. of Subjects	Reference
Circles									
17.	.82	$\rho > .99$	magnitude estimation (physical area)	179	11.0 or 17.4 always present	Low	10 or 100	8	MacMillan, Moschetto et al. (1974) cont. experiment 1
18.	.81	$\rho > .99$	magnitude production (physical area)	3075	48.0 or 75.5 always present	Low	10 or 100	8	"
19.	.59	$\rho > .99$	magnitude estimation (apparent size)	179	11.0 or 17.4 presented once	Low	10 or 100	6	"
20.	.58	$\rho > .99$	magnitude production (apparent size)	3075	48.0 or 75.5 presented once	Low	10 or 100	6	"
21.	.65	$\rho > .99$	magnitude estimation (physical size)	179	11.0 or 17.4 presented once	Low	10 or 100	8	"
22.	.66	$\rho > .99$	magnitude production (physical size)	3075	48.0 or 75.5 presented once	Low	10 or 100	8	"
23.	.84	$\rho > .99$	magnitude estimation (physical area)	85	1 always present	Low	1 or 10	8	MacMillan, Moschetto et al. (1974) cont. experiment 2

[tt] Ratio setting is a modification of fractionation. Ratio estimation is a modification of the method of constant sums

* The different measures of variability used by investigators are: σxy - sample std dev from regression (log-log plot), σ - std dev for distribution of individual subject exponents, $\sigma x^2 y$ - residual variance about regression (log-log plot), R - range of individual subject exponents, Rw - width of range of individual subject exponents, ρ - correlation between log ϕ and logψ.

TABLE 7
(Continued)

POWER LAW EXPONENTS FOR AREA OF VARIOUS FIGURES

No.	Exponent	Variability* Measure	Method	Stimulus Ratio Max Area/Min Area	Standard Std Area/ Min Area	Location of Std in Range	Modulus	No. of Subjects	Reference
Circles									
24.	.97	$\rho > .99$	magnitude estimation (physical area)	85	13.8 always present	Middle	10 or 100	8	MacMillan, Moschetto et al. (1974) cont. experiment 2
25.	.80	$\rho > .99$	magnitude estimation (physical area)	85	85 always present	High	100 or 1000	8	"
26.	.75	$\rho > .99$	magnitude estimation (physical area)	85	1 presented once	Low	1 or 10	8	"
27.	.70	$\rho > .99$	magnitude estimation (physical area)	85	13.8 presented once	Middle	10 or 100	8	"
28.	.71	$\rho > .99$	magnitude estimation (physical area)	85	85 presented once	High	100 or 1000	8	"

---

tt Ratio setting is a modification of fractionation. Ratio estimation is a modification of the method of constant sums

* The different measures of variability used by investigators are: σxy - sample std dev from regression (log-log plot), σ - std dev for distribution of individual subject exponents, $\sigma x^2 y$ - residual variance about regression (log-log plot), R - range of individual subject exponents, Rw - width of range of individual subject exponents, ρ - correlation between log ϕ & log ψ.

3,7

TABLE

POWER LAW EXPONENTS FOR VOLUME OF VARIOUS SOLIDS

No.	Exponent	Variability* Measure	Method	Stimulus Range (Max Vol/Min Vol)	Standard (Std Vol/ Min Vol)	Location of Std in Range	Modulus	No. of Subjects	Reference
Cubes									
1.	1.01	-	ratio estimation	9.5	N.A.	N.A.	N.A.	10	Ekman, Junge (1961)
2.	.07	σxy = .02	magnitude estimation	1000	78	mid	10	10	Teightsoonian (1965)
3.	.72	σxy = .02	magnitude estimation	145	11.4	mid	10	10	"
Octahedrones									
4.	.65	σxy = .04	magnitude estimation	1060	46	mid	10	10	"
5.	.74	σxy = .04	magnitude estimation	78	8	mid	10	10	"

* The different measures of variability used by investigators are: σxy - sample std dev from regression (log-log plot), σ - std dev for distribution of individual subject exponents, $\sigma x^2 y$ - residual variance about regression (log-log plot), R - range of individual subject exponents, Rw - width of range of individual subject exponents.

** N.A. - parameter is Not Applicable

3.9

TABLE 10-

POWER LAW EXPONENTS FOR APPARENT SIZE OF PERSPECTIVE DRAWINGS

No.	Volume (perspective drawings)	Exponent	Variability* Measure	Method	Stimulus Range (Max Vol/Min Vol)	Standard (Std Vol/ Min Vol)	Location of Std in Range	Modulus	No. of Subjects	Reference
1.	cube	.79	-	ratio estimation	9.5	N.A.	N.A.	N.A.	10	Ekman, Junge (1961) experiment 1
2.	cube	.75	-	magnitude estimation	3500	100	mid	100	12	Ekman, Junge (1961) experiment 3
3.	sphere	.74	-	magnitude estimation	3500	100	mid	100	12	"
4.	various cubes and spheres	.69	S.E. = .05	magnitude estimation	3500	60,600	low, mid	100, 1000	186	Ekman, Lindman, William-Olsson (1961) experiments 1 & 2
5.	cubes and spheres with surface texture	.59	-	magnitude estimation	3500	60	low	100	99	Ekman, Lindman, William, Olsson (1961) experiment 3

* The different measures of variability used by investigators are: σxy - sample std dev from regression (log-log plot), σ - std dev for distribution of individual subject exponents, $\sigma x^{2} y$ - residual variance about regression (log-log plot), R - range of individual subject exponents, Rw - width of range of individual subject exponents.

** N.A. - parameter is Not Applicable

349,9

3,10

TABLE ~~11~~

POWER LAW EXPONENTS FOR PROPORTION AND NUMEROUSNESS

No.	Exponent	Variability* Measure	Method	Stimulus Ratio	Standard	Location of Std in Range	Modulus	No. of Subjects	Reference
Proportion									
1.	.97	σ = .38	magnitude estimation	80 elements (dots & lines) 5/80 - 75/80	40/80	mid	10	30	Rule (1968)
Numerousness									
2.	1.34	-	fractionation	2-180 dots	N.A.	N.A.	N.A.	5	Stevens, S.S. (1957) based on data by Taves (1941)
3.	1.03	σ = .23	magnitude estimation	9- 82 dots	27	mid	10	36	Rule (1966)
4.	.72	-	magnitude estimation	25-200 dots	none used	N.A.	N.A.	30	Krueger (1972) experiment 1
5.	.78	-	magnitude estimation	25-200 dots	none used	N.A.	N.A.	32	Krueger (1972) experiment 2
6.	.77	-	magnitude estimation	25-400 X's	none used	N.A.	N.A.	32	Krueger (1972) experiment 3
7.	.93	-	magnitude production	25-200 X's	none used	N.A.	N.A.	32	Krueger (1972) experiment 4

tt Ratio setting is a modification of fractionation. Ratio estimation is a modification of the method of constant sums.

* The different measures of variability used by investigators are: σxy - sample std dev from regression (log-log plot), σ - std dev for distribution of individual subject exponents, $\sigma x^2 y$ - residual variance about regression (log-log plot), R - range of individual subject exponents, Rw - width of range of individual subject exponents.

** N.A. - parameter is Not Applicable

[illegible] 3,10

Table ~~8~~ 3.11

Exponents For Saturation of Surface Colors

Hue	Wave Length (nm)	Luminance Factor (% reflectance)	Artificial Light (69.7 db)		Daylight	
			4°	0.7°	4°	0.7°
Bluish purple	425	18.6	1.77			
Purplish blue	462	16.0	1.44			
Blue	473	19.5	1.50	1.97	1.11	1.94
Greenish blue	481	17.5	1.97			
Blue green	496	18.6	2.00			
Green	521	20.5	1.97			
Yellowish green	556	27.5	2.84	3.09	2.46	2.86
Greenish yellow	573	47.9	4.06			
Yellow	577	53.4	3.58	2.85	4.01	2.84
Orange	588	17.6	2.60	2.96	2.74	3.01
Orange pink	604	13.5	2.17			
Pink	614	15.1	2.26			
Pinkish red	630	21.5	2.24	1.73	1.60	1.84
Purplish pink	499	18.3	2.39			
Reddish purple	562	19.7	1.96			

From Guirao & de Mattiello (197~~0~~4)

352 Table. 3.11

Figure 2

LABELS VIOLATING BOUNDARIES OF IDENTIFICATION ABILITY

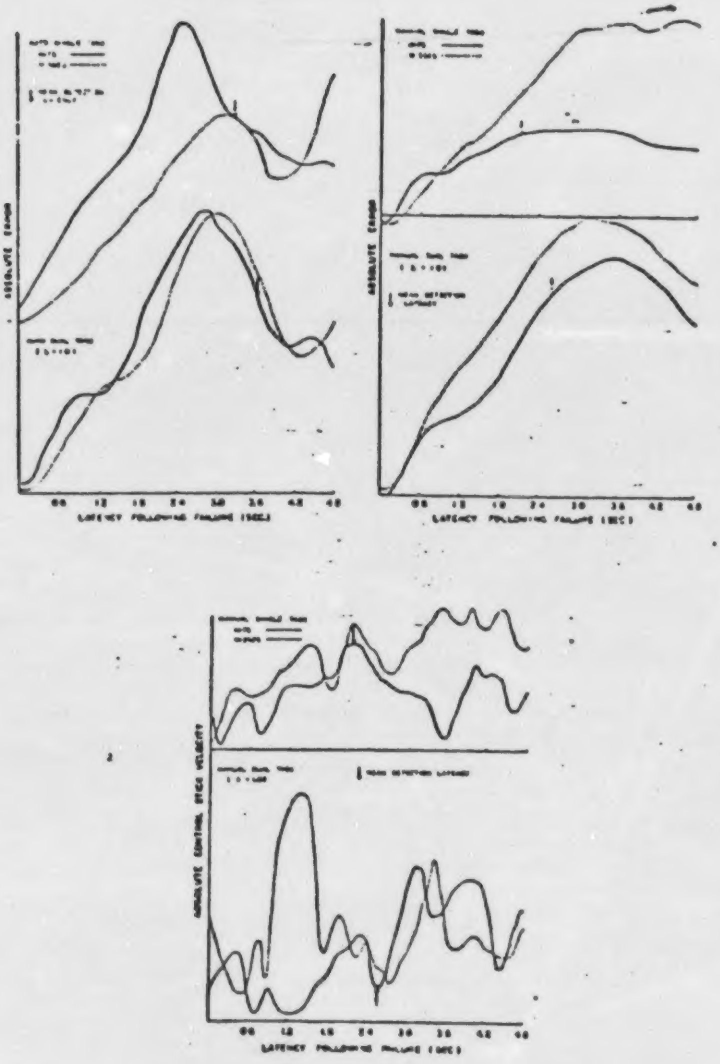


Fig 3,1

3,12
Table ~~7~~

Exponents For Lightness of Surface Colors

Hue	Colormetric Purity	Range of Luminance Factor (%)	Exponent	
			Daylight	Artificial Light
Gray	-	.4-80	1.07	.92
Blue (470 nm)	.03	5-59	.84	.67
	.05	4-40	.73	.77
	.07	8-25	1.03	1.19
	.09	7-30	.86	.80
	.16	7-19	.78	.65
Green (553 nm)	.23	9-63	.90	.80
	.26	9-61	1.04	1.03
	.34	16-41	.92	.76
	.39	13-47	.25	.69
	.43	16-33	.73	.70
Yellow (574 nm)	.36	9-68	1.99	.48
	.47	8-38	1.02	.72
	.57	19-70	1.04	.76
	.69	20-72	1.90	.52
	.76	28-75	.94	.50
Red (622 nm)	.07	17-43	1.12	.92
	.11	6-55	.88	.86
	.16	6-41	.90	.90
	.23	7-28	1.04	.96
	.51	3-17	.68	.62

From de Mattiello & Guirao (1974)

35

TAble 3,17

Figure 3

LUMINANCE LEVELS FOR A NUMBER OF COMMONLY EXPERIENCED CONDITIONS

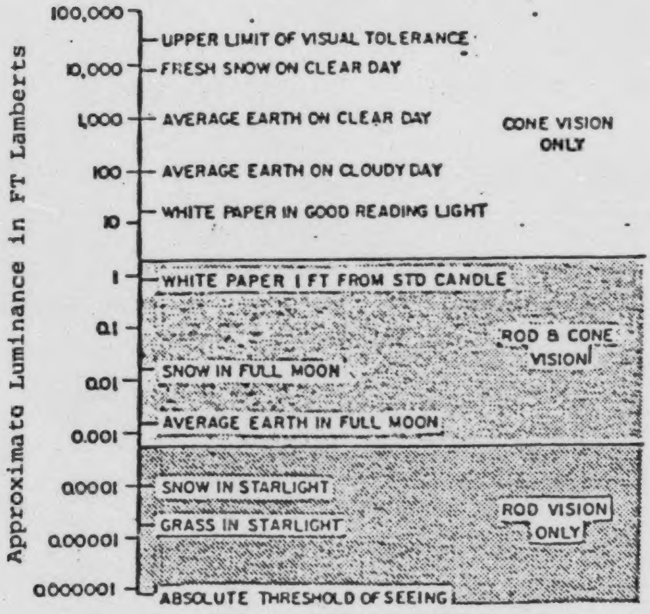


Fig 3.2

3,3

Figure 4

VISUAL ACUITY AS A FUNCTION OF BACKGROUND LUMINANCE
(from Moon and Spencer, 1944)

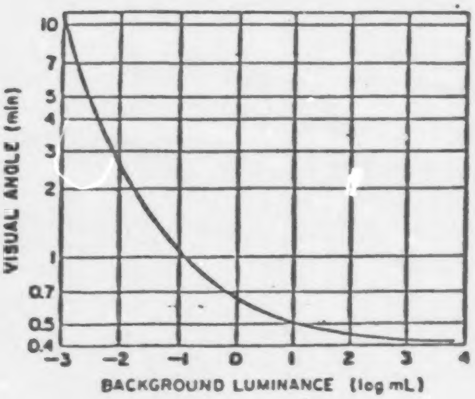


Fig 3,3

Figure ~~5~~ 3.4.

CONTRAST SENSITIVITY FOR SQUARE WAVE GRATINGS (□) AND SINE WAVE GRATINGS (○). THE LUMINANCE OF GRATINGS FOR UPPER PAIR OF CURVES WAS 500 c/m² AND 0.05 c/m² FOR LOWER PAIR OF CURVES. (from Campbell and Robson, 1968)

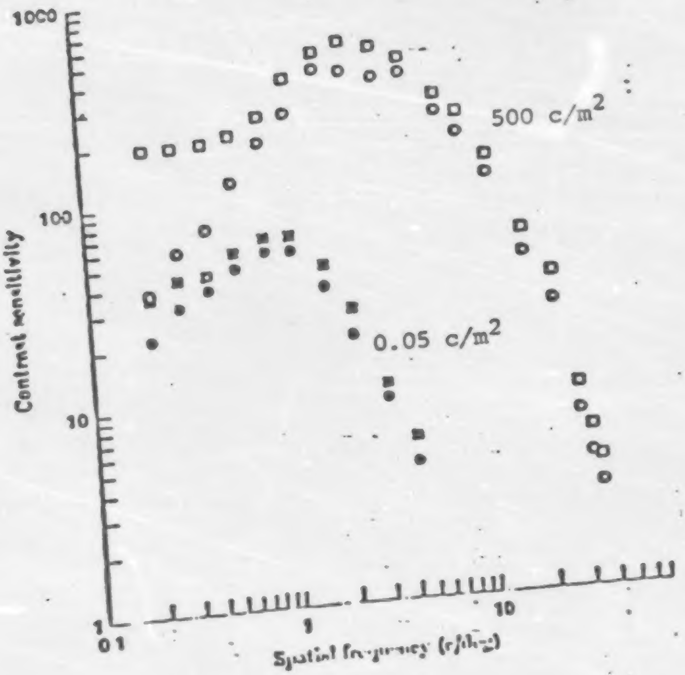


Fig 3.4

Figure ~~6~~ 3.5

THE EFFECTS OF CONTRAST ON ACUITY (from ,)

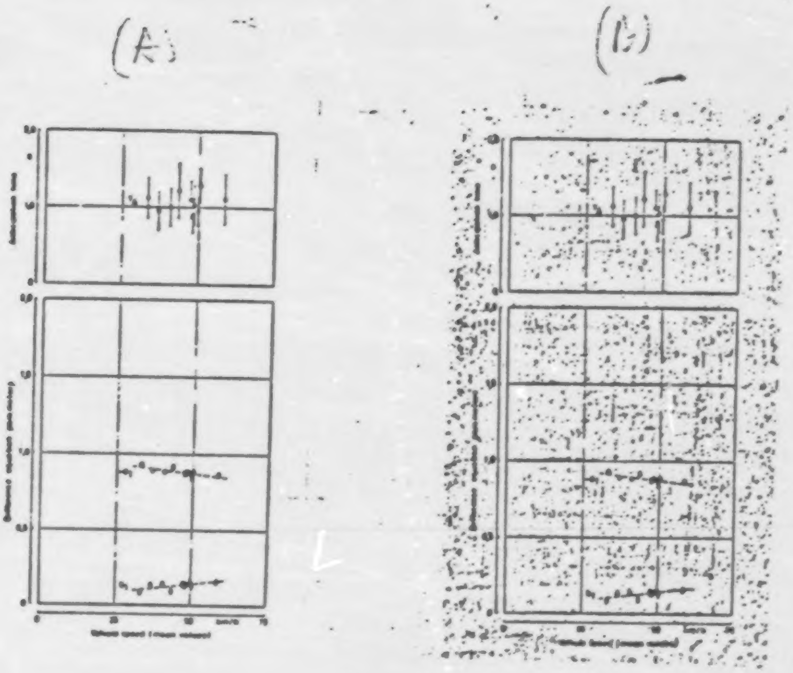
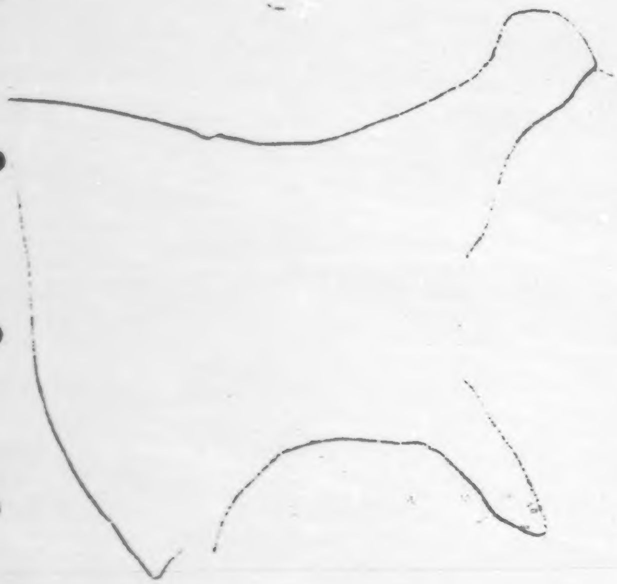


fig 3.5

(a)



(b)

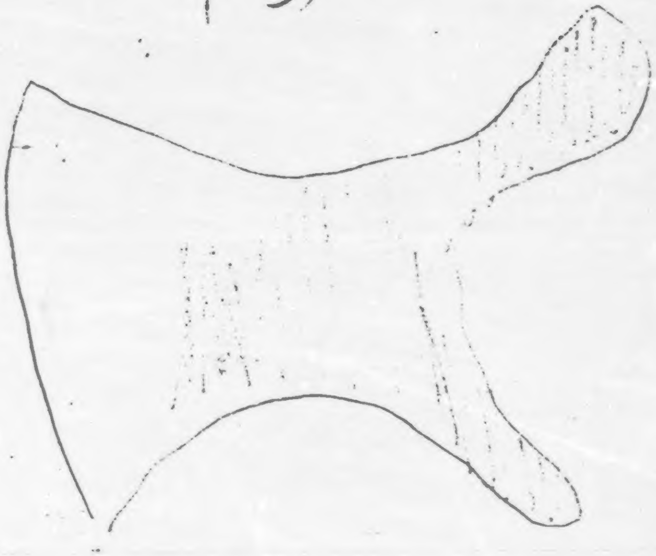
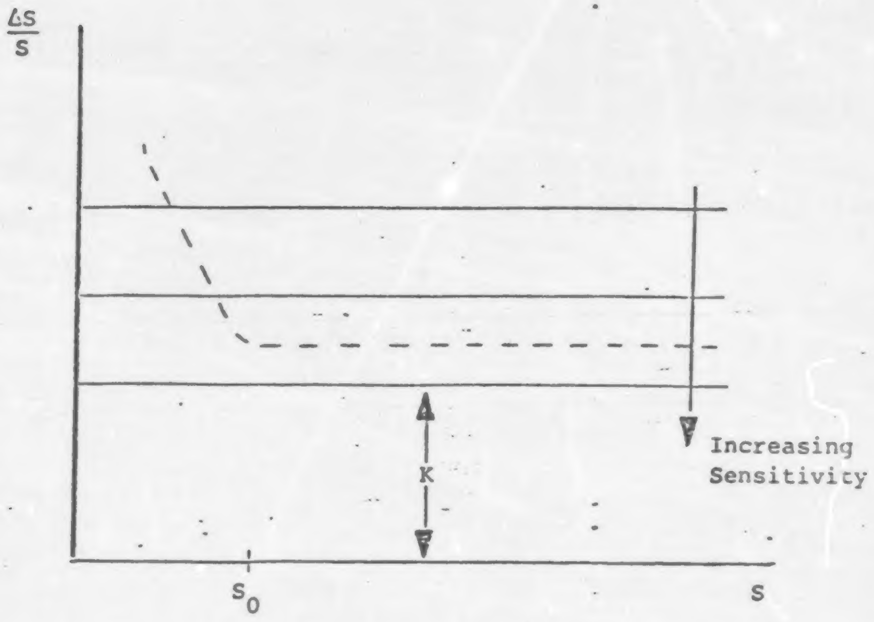


fig 3,6

Figure 8

GRAPH SHOWING THEORETICAL (solid lines) AND EMPIRICAL (dashed line) FORM OF WEBER's LAW



36J

(Fig 3,7

Figure ~~9a~~ 2.8

EXPENDITURES
(Less than one JND)

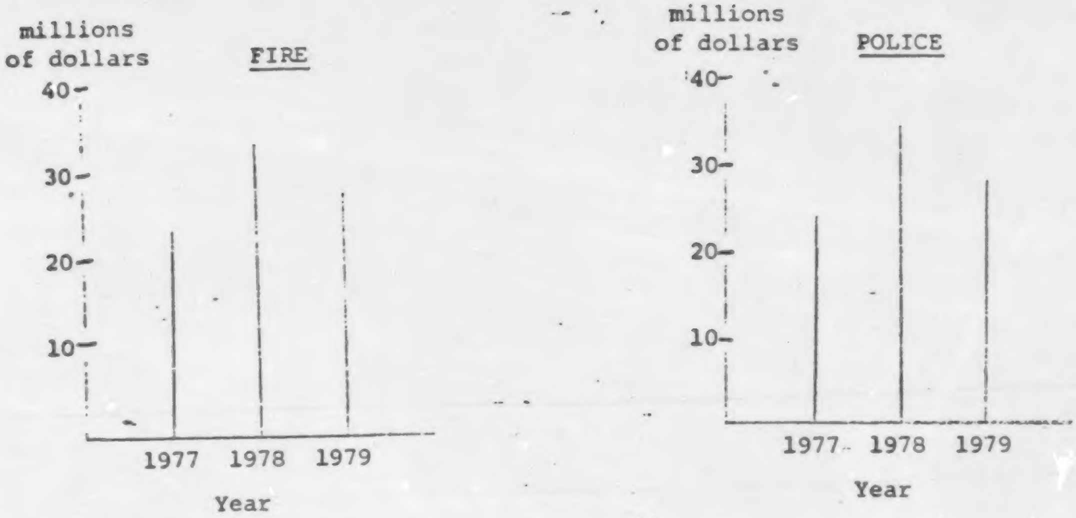


Fig 3, 8a

3.5b

Figure 9b

EXPENDITURES
(Greater than one JND)

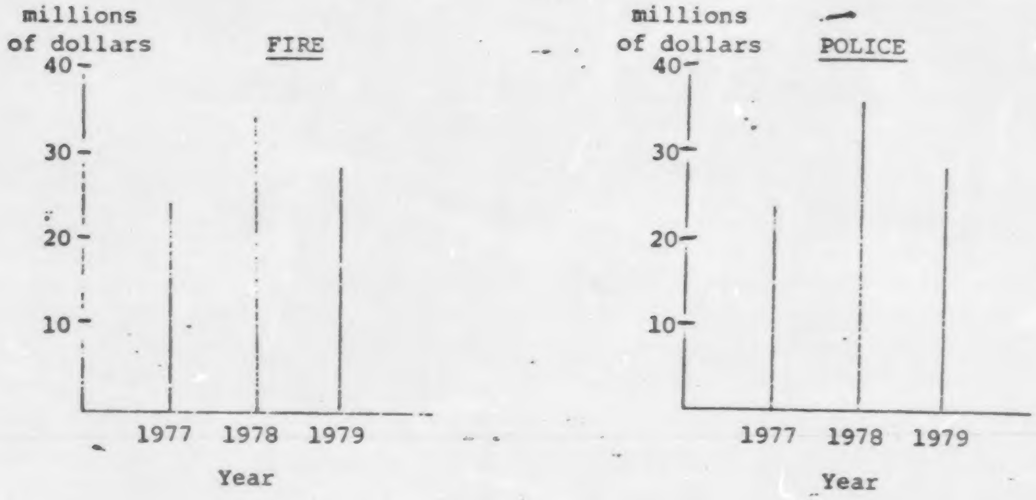


Fig 3.8b

Figure 10a

SIMPLE NEAREST-NEIGHBOR SINGLE-LINE AGGLOMERATIVE CLUSTERING (from Marriott, 1974)

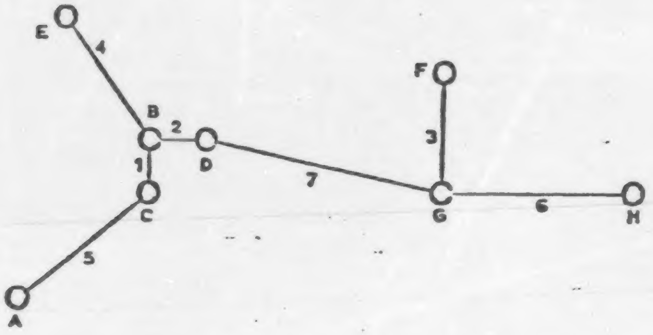

E
4
B
2
F
3
1
D
7
C
G
6
H
5
A

Fig 3, 9 a

Figure 10b

SINGLE LINK CLUSTERING WITH AC LINK
REDRAWN TO HORIZONTAL PLANE

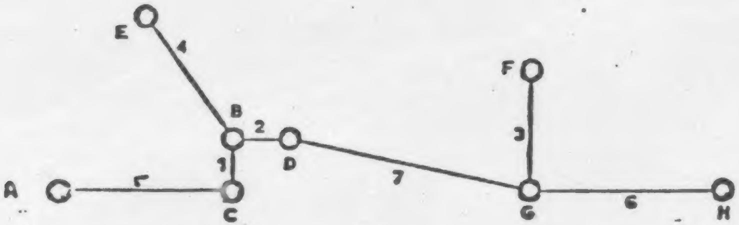
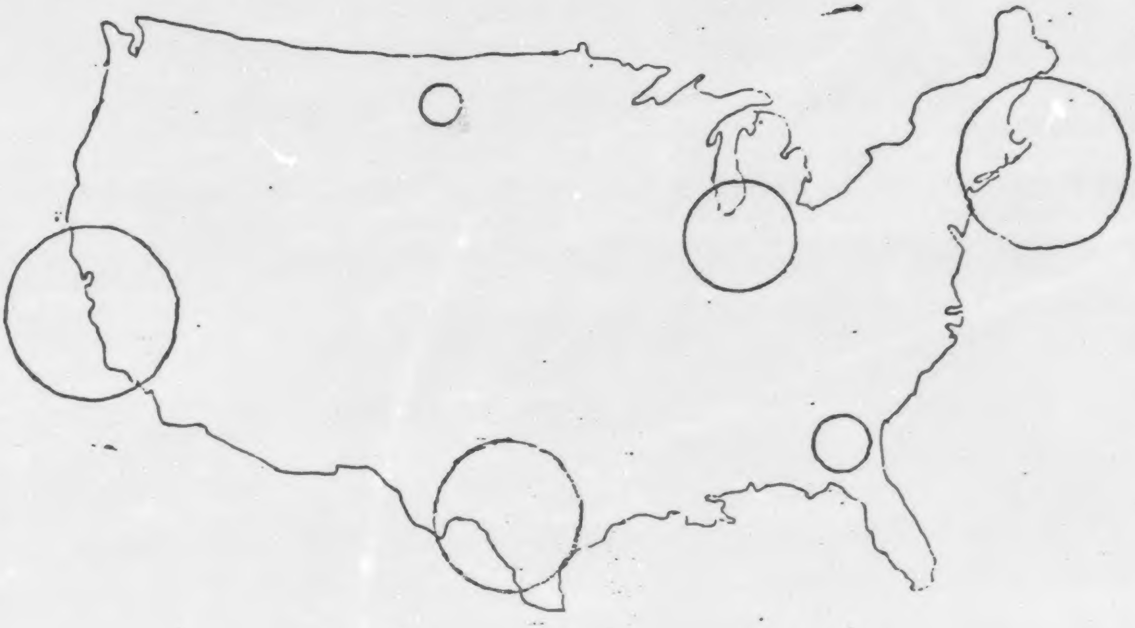

E
4
F
B
2
3
1
D
7
A
5
C
G
6
H

Fig 3.9 b

3,10a

Figure 11a

TECHNOLOGY MANPOWER FOR DIFFERENT REGIONS OF U.S.
(Less than one JND)



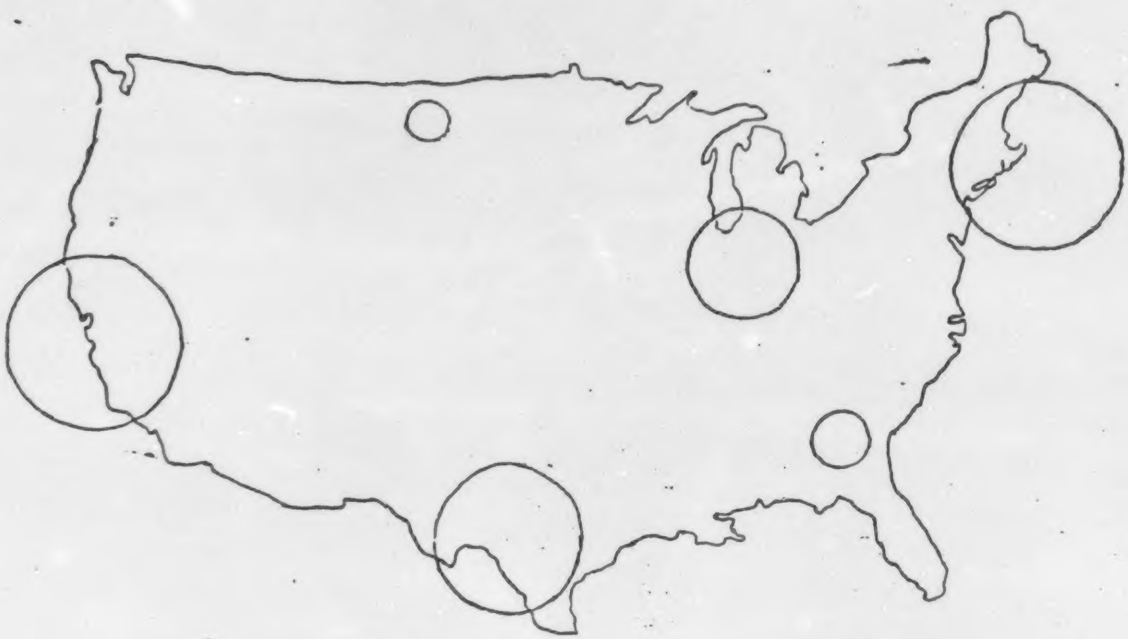
3, 10 a

Fig. 3, 10 a

3.12b

Figure 11b

TECHNOLOGY MANPOWER FOR DIFFERENT REGIONS OF U.S.
(Greater than one JND)



3,11

Figure 12

RELATION OF HUE TO WAVELENGTH

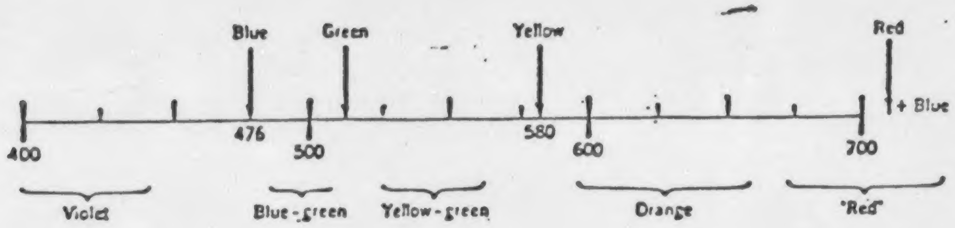

Blue
Green
Yellow
Red
+ Blue
400
476
500
580
600
700
Violet
Blue-green
Yellow-green
Orange
"Red"

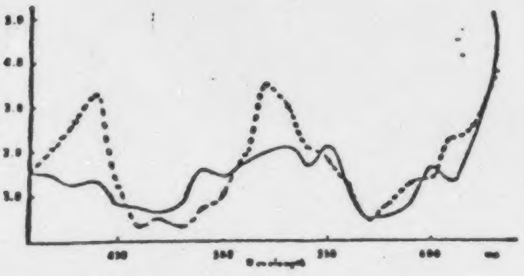
Fig 3,11

3,12

Figure 13

MEAN JND's AND STANDARD DEVIATIONS FROM 410-63. m..
(from Siegel, 1964)

●----● JND's
●——● Standard Deviations



410

(will be redrawn!)

Fig 3,12

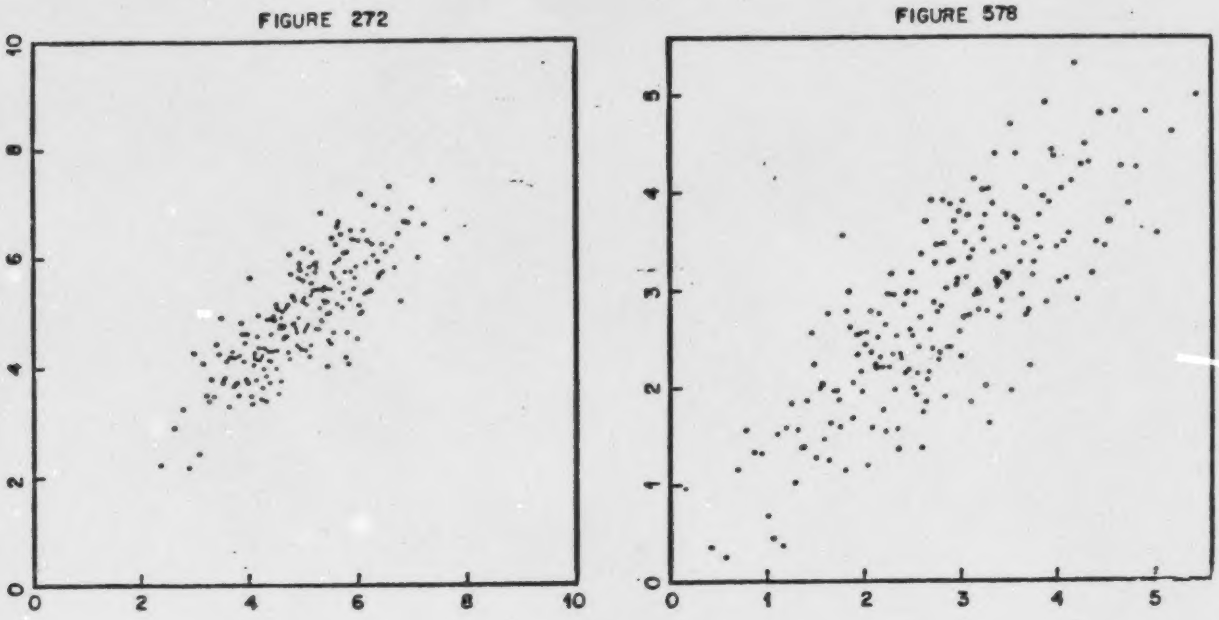
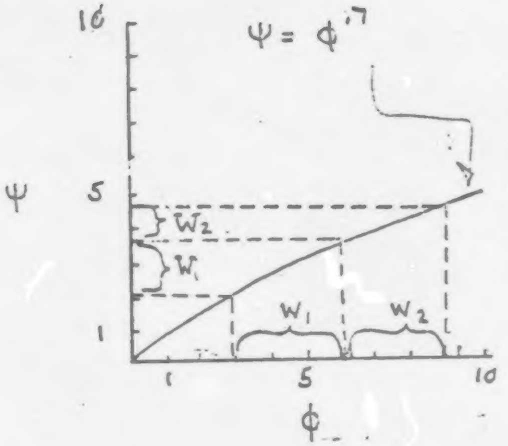


Figure 1. Reductions of two scatterplots used in the three types of experiments. The left panel is point-cloud size 2 and the right panel is point-cloud size 4.

Fig 3.126

(A)



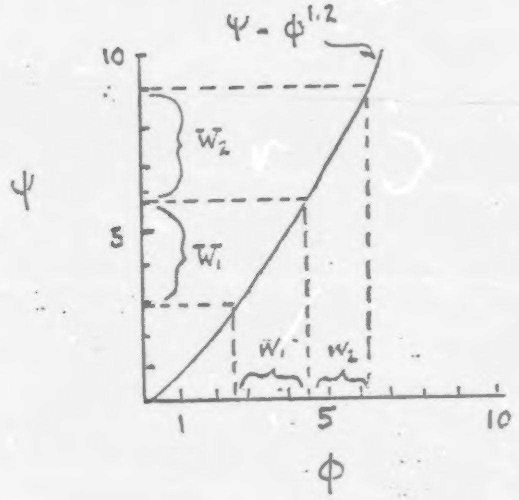
Psychophysical Function (b = .7)

$W_1 = W_2$

$W_1 > W_2$

FIGURE 14a

(B)



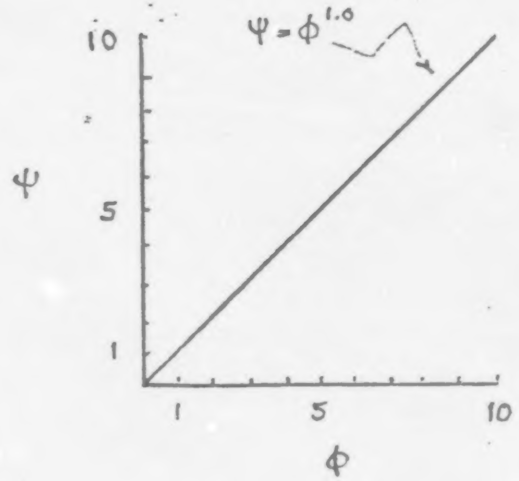
Psychophysical Function (b = 1.2)

$W_1 > W_2$

$W_1 = W_2$

FIGURE 14b

(C)



Psychophysical Function (b = 1.0)

FIGURE 14c

Fig 3.13

Lightness as a Function of
Reflectance For Red (622nm)

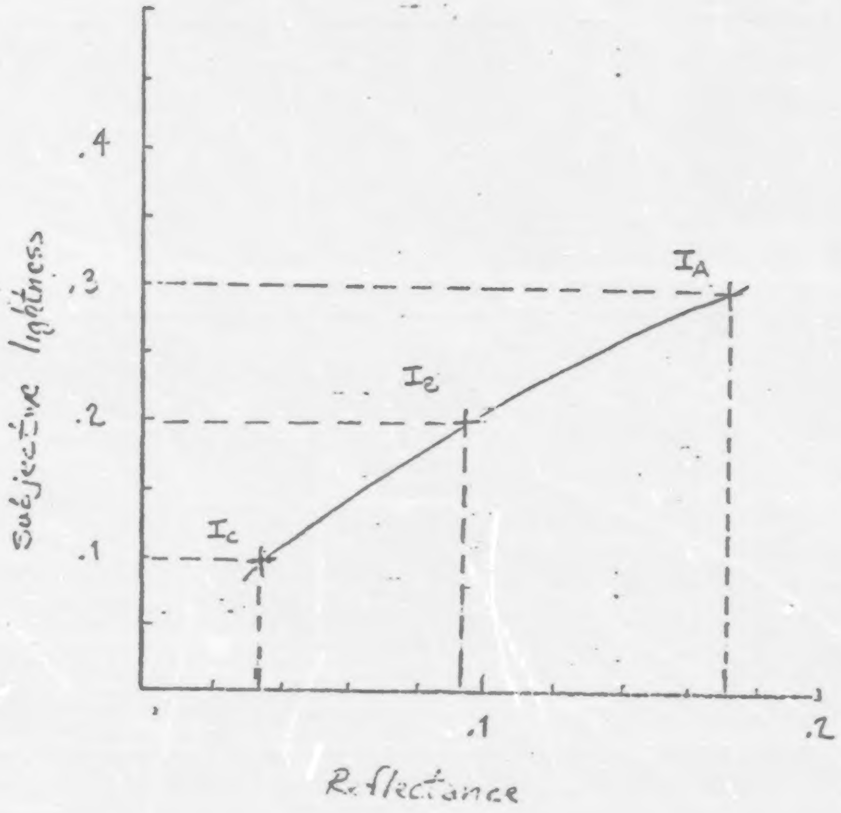
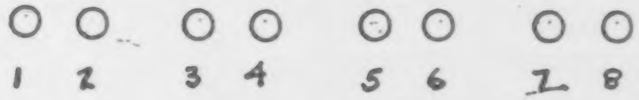
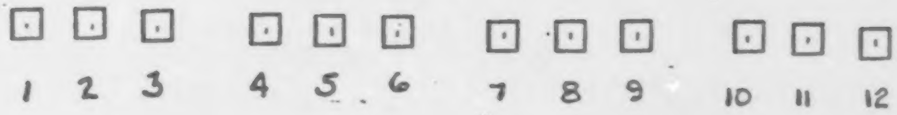


FIGURE 15

fig 3,14



3.14
FIGURE 1b-a


1
2
3
4
5
6
7
8
9
10
11
12

3.14
FIGURE 1b-b

fig 3.14

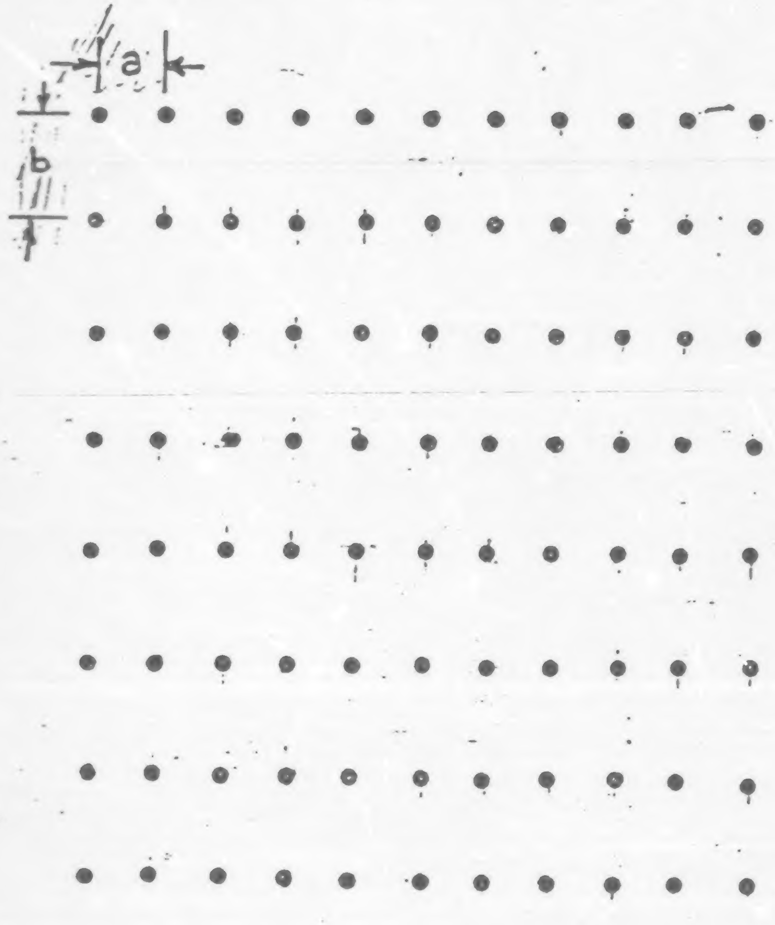


FIGURE 17

fig 3.15

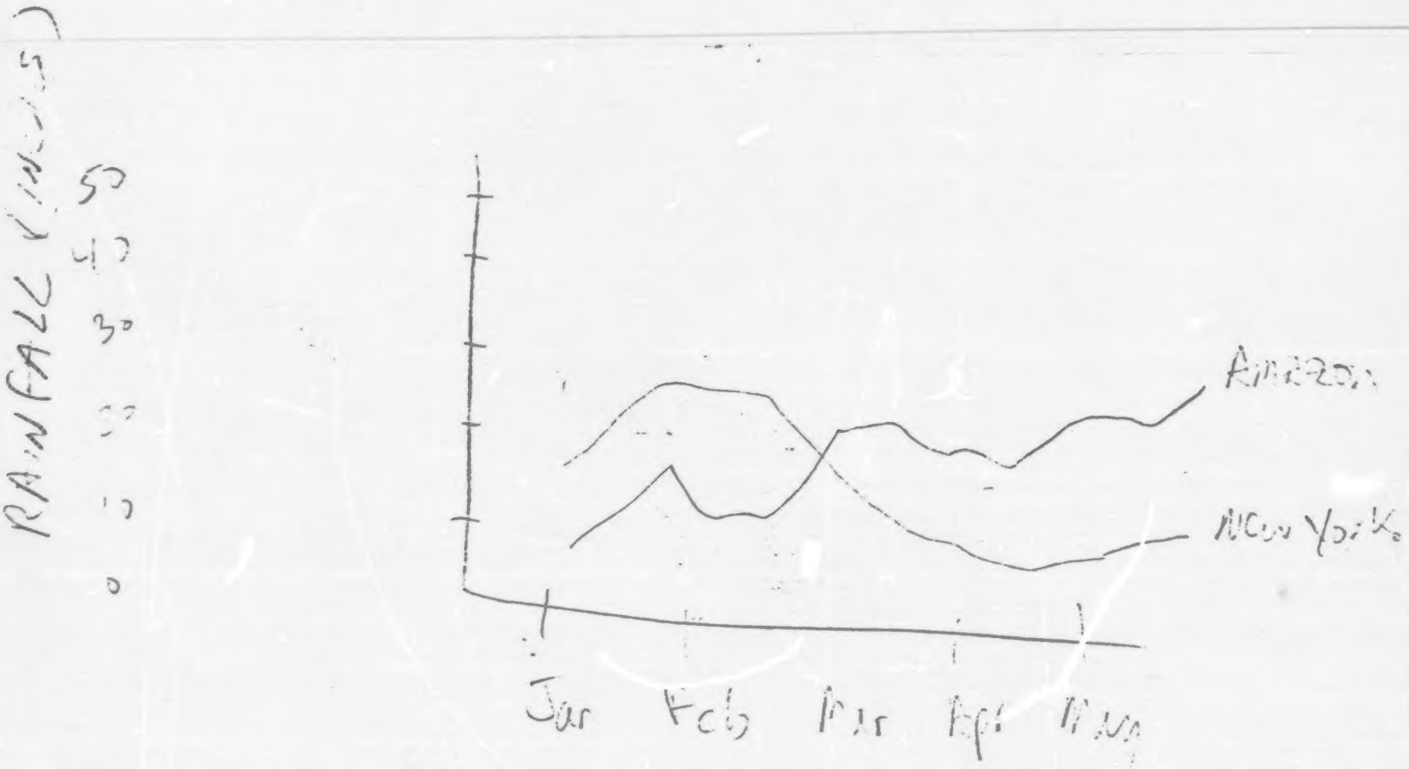
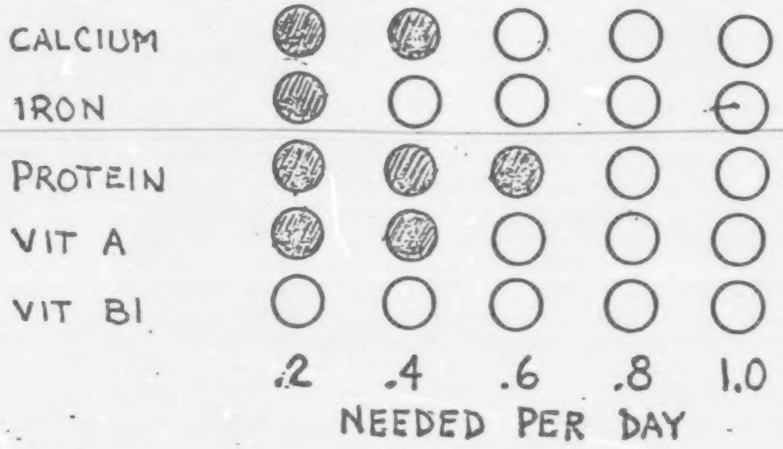

RAINFALL (INCHES)
50
40
30
20
10
0
Amazon
New York
Jan
Feb
Mar
Apr
May

Fig 3.16

(a)



3.17a

FIGURE 79a

b)

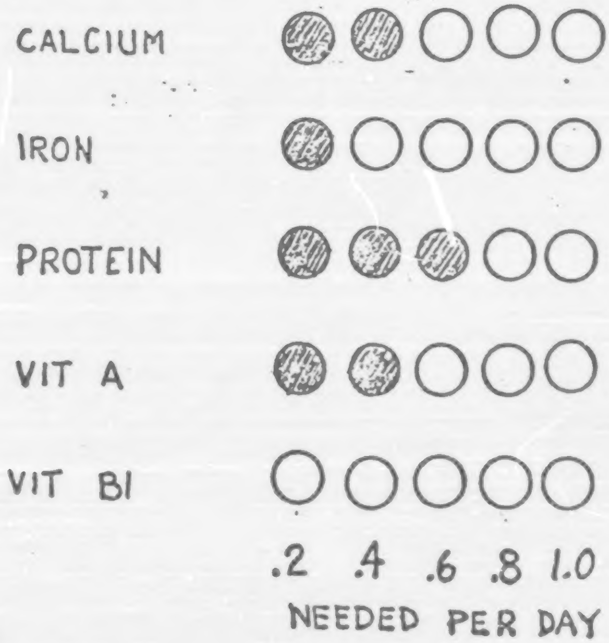


fig 3.17b

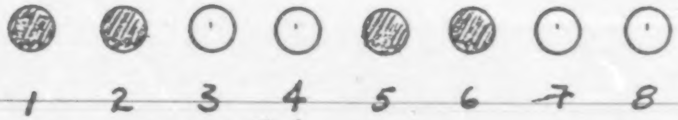


FIGURE 20a? 18a

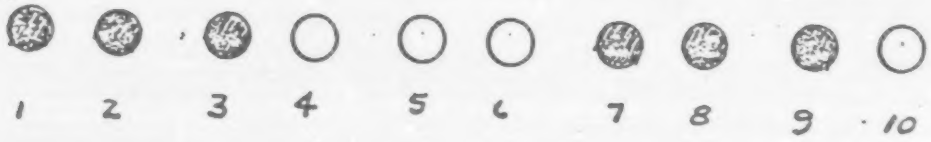

1
2
3
4
5
6
7
8
9
10

FIGURE 20b 3.18

fig 3.18

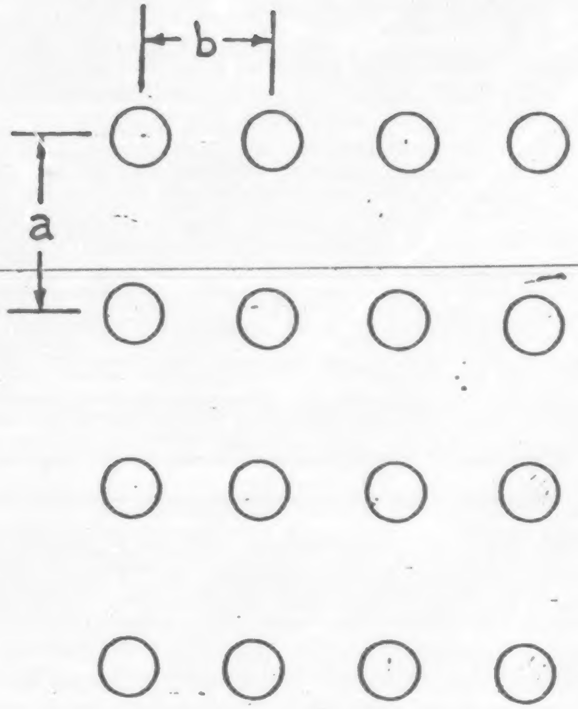
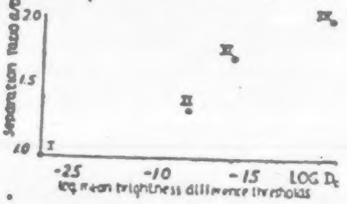


FIGURE 21a. Hochberg & Silverstein (1956) asked subjects to [illegible] row distance (a) with column distance (b), measuring grouping based on proximity vs. brightness, shape, etc.



Brightness differences needed to overcome proximity in grouping

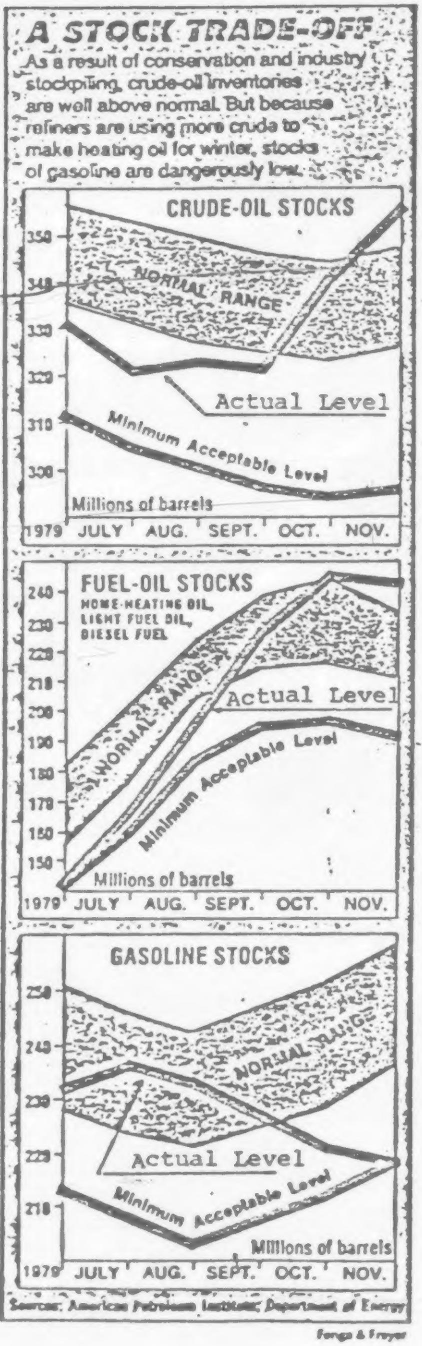
Wi[illegible]

[illegible]

~~FIGURE 21b~~

Fig 3.19

16

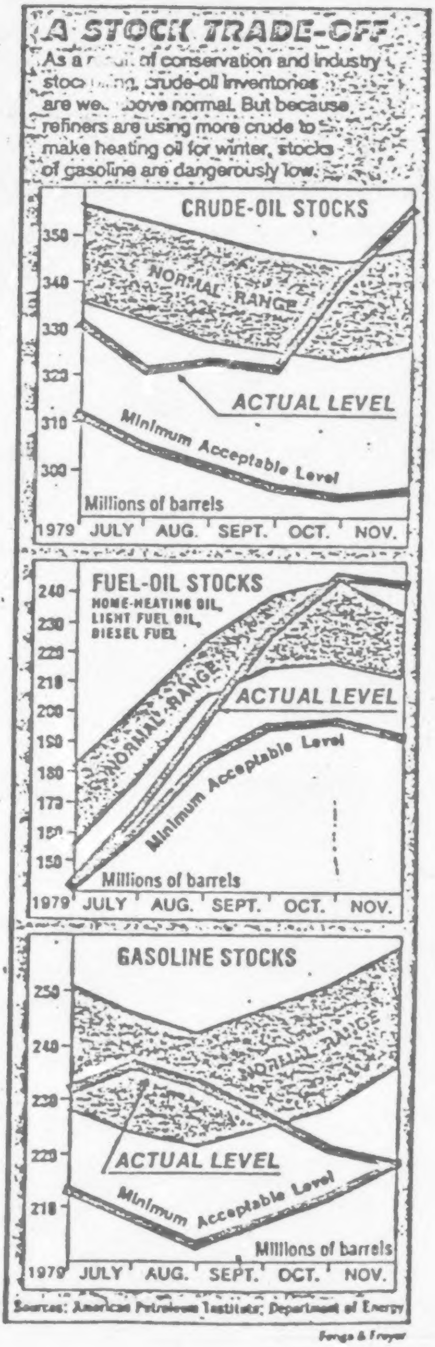


Sources: American Petroleum Institute; Department of Energy

Fenga & Freyer

FIGURE 22

Fig 3.20



Sources: American Petroleum Institute; Department of Energy

Fenga & Freyer

FIGURE 23

Fig 3.21

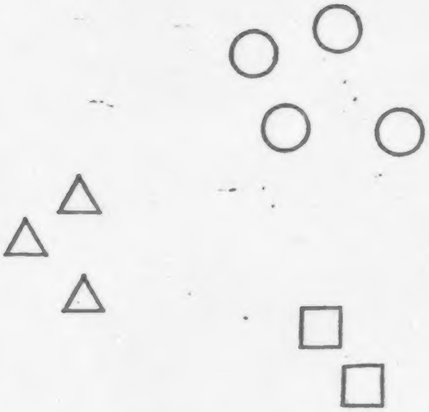


FIGURE ~~24~~ 3,22

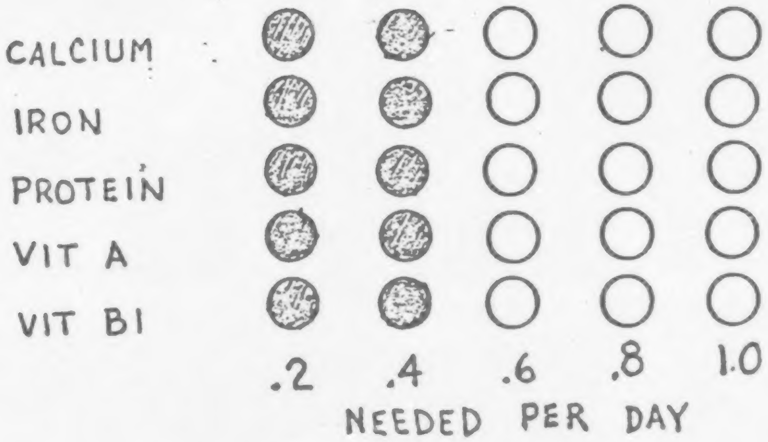


FIGURE ~~25~~ 3.23

Figs 3,23

1

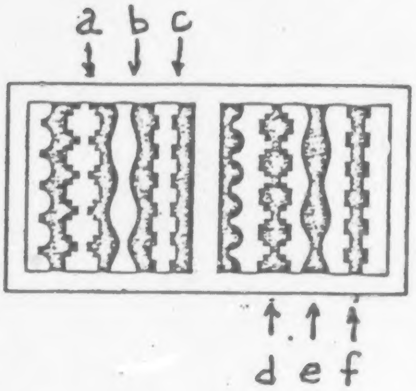
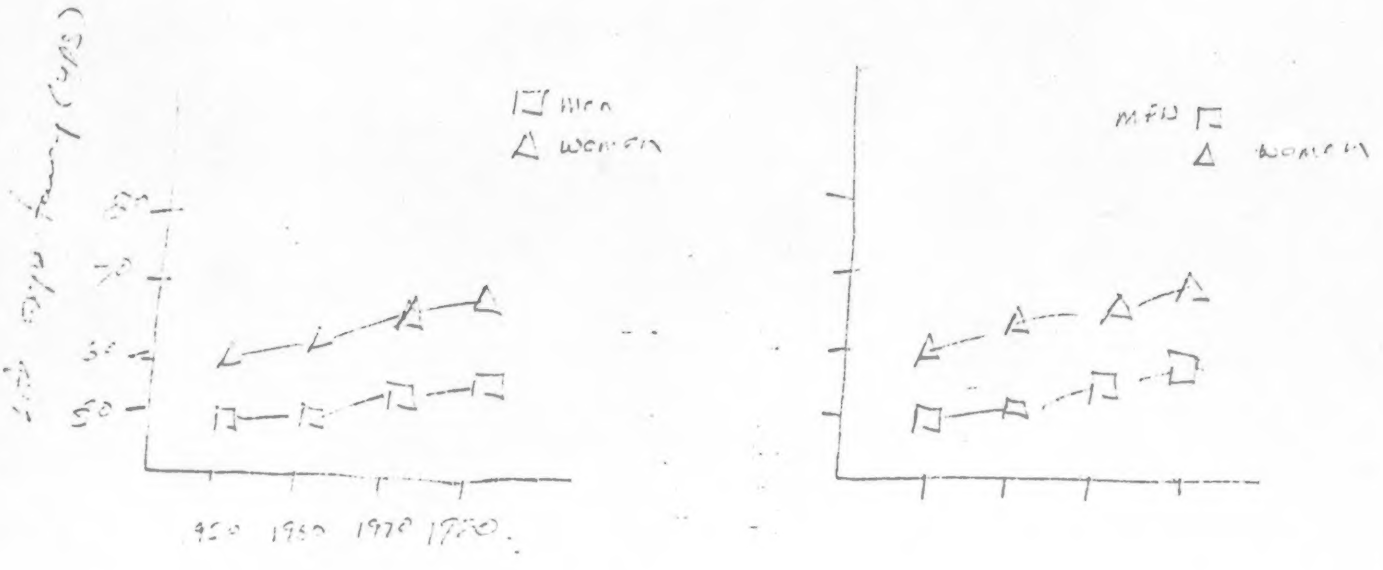


FIGURE ~~26~~ 3:24

380

fig 3, 24


Life expectancy (yrs)
80
70
60
50
1950 1960 1970 1980.
□ men
△ women
MEN □
△ WOMEN

381 Fig 3.25

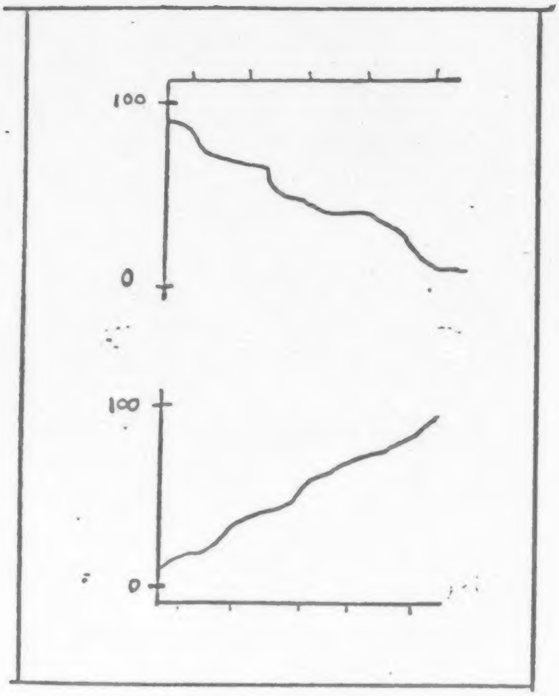


FIGURE 28-a

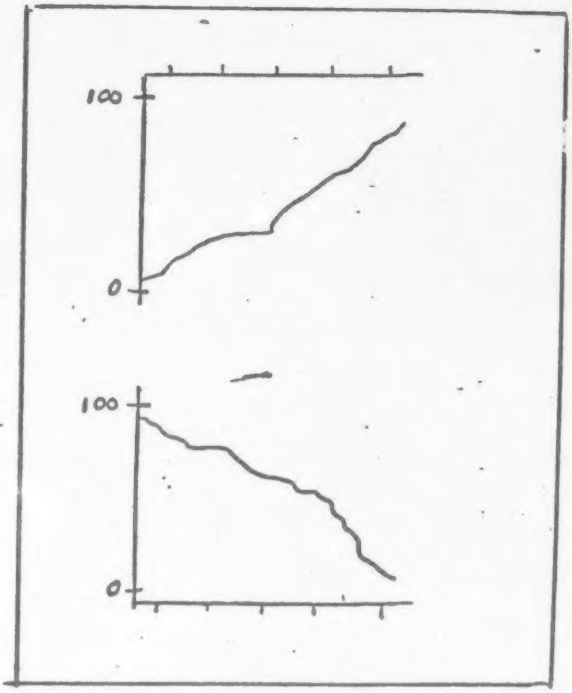


FIGURE 28-b Fig 3.26

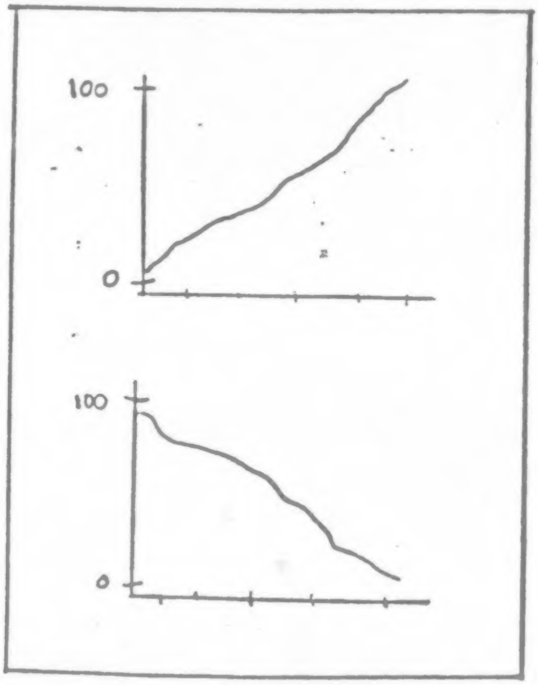


FIGURE 29-a

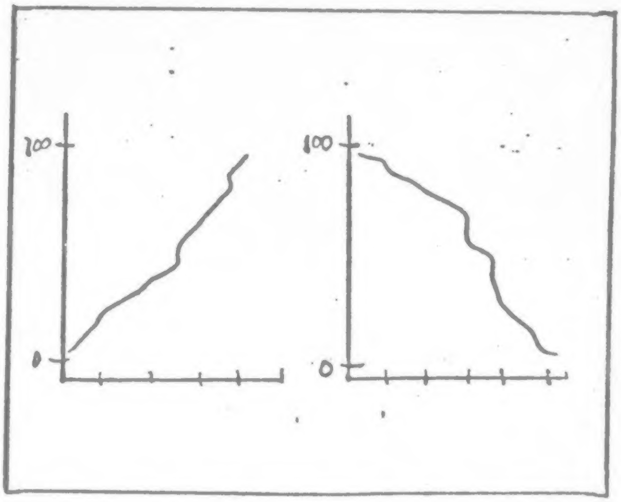


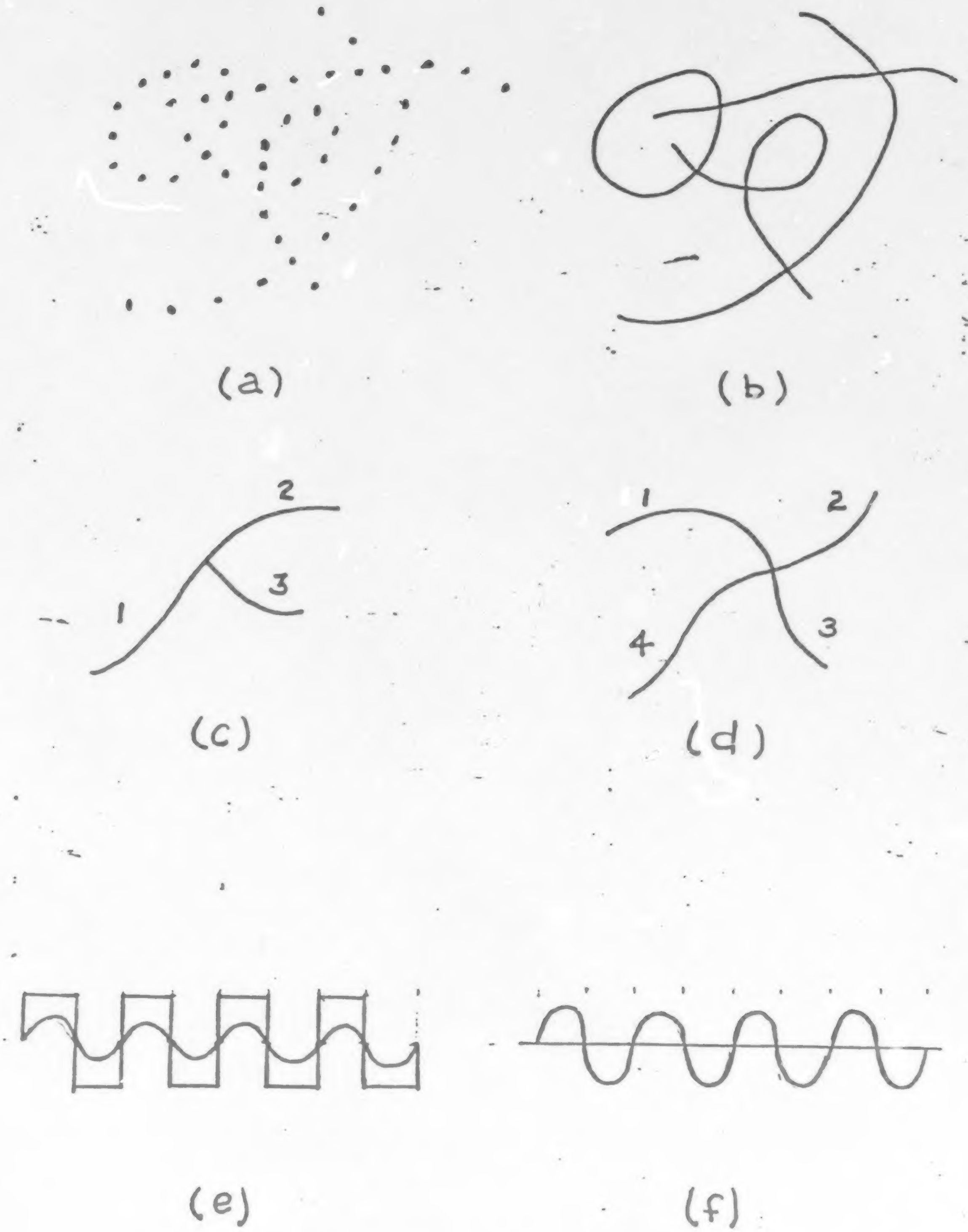
FIGURE 29-b

fig 3.27

3.28

22

FIGURE 30

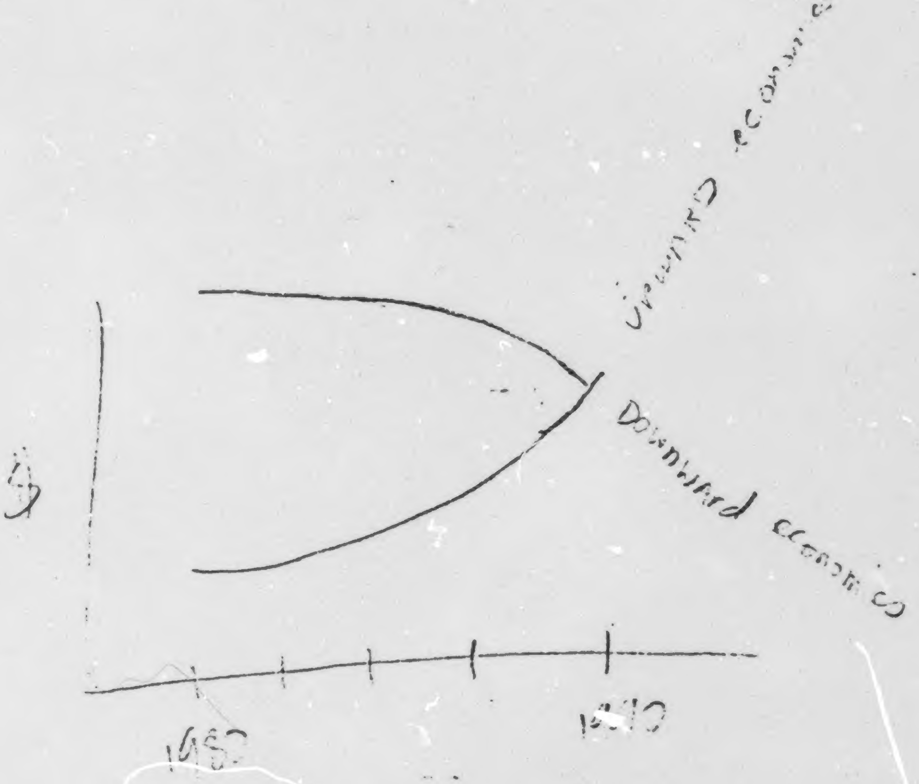


(a) (b)

(c) (d)

(e) (f)

383 F. 13.28



3,29

fig 3,29

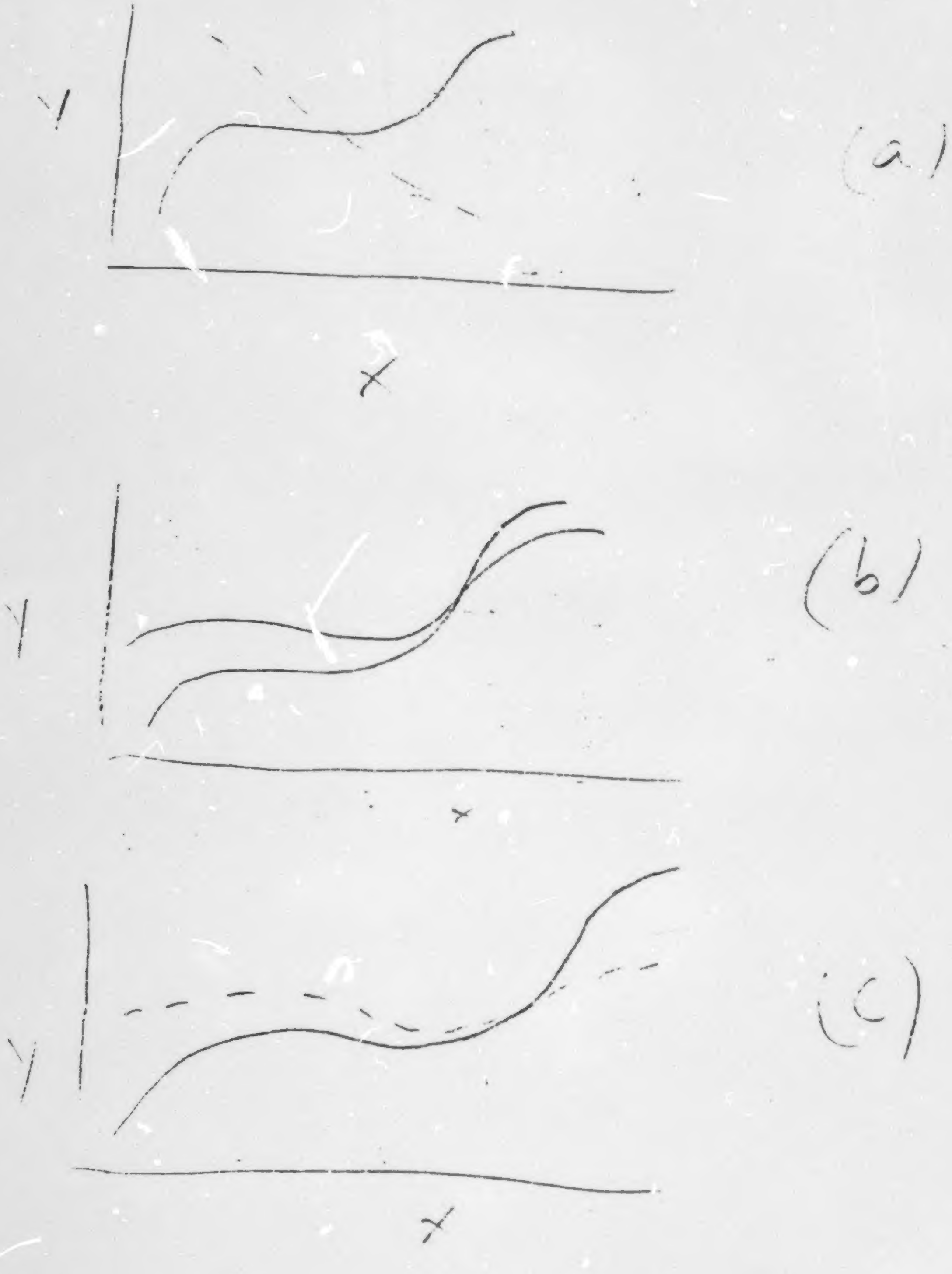


Fig 3.30

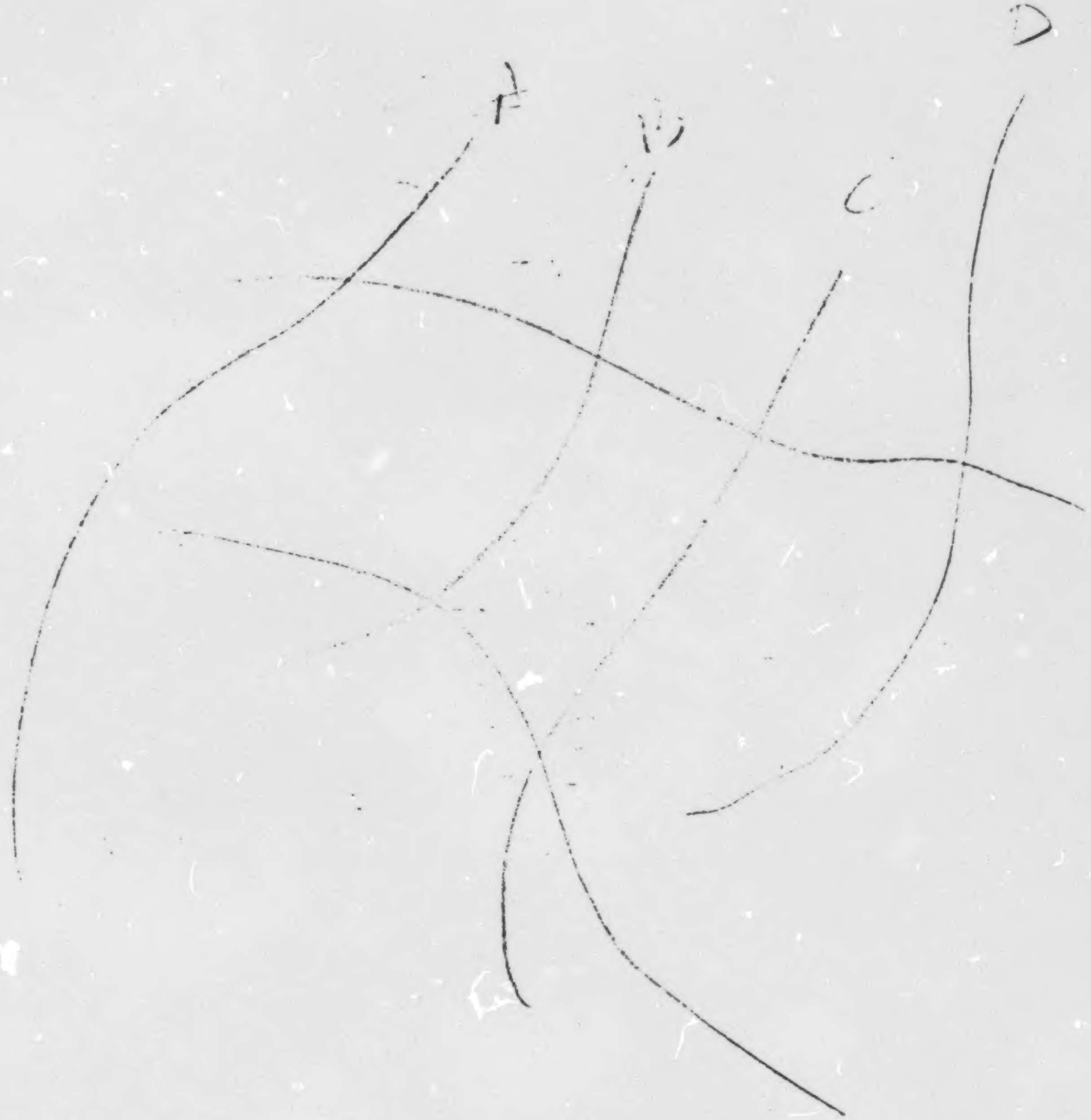


fig 3.31

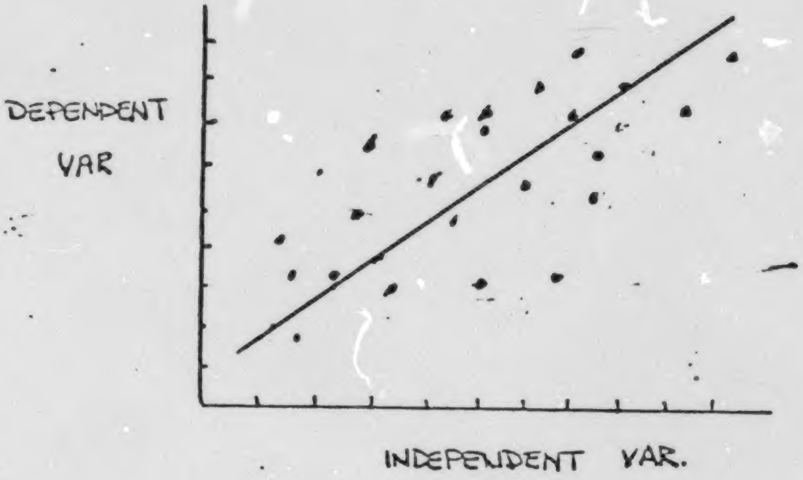


FIGURE ~~34~~ Fig 3.32

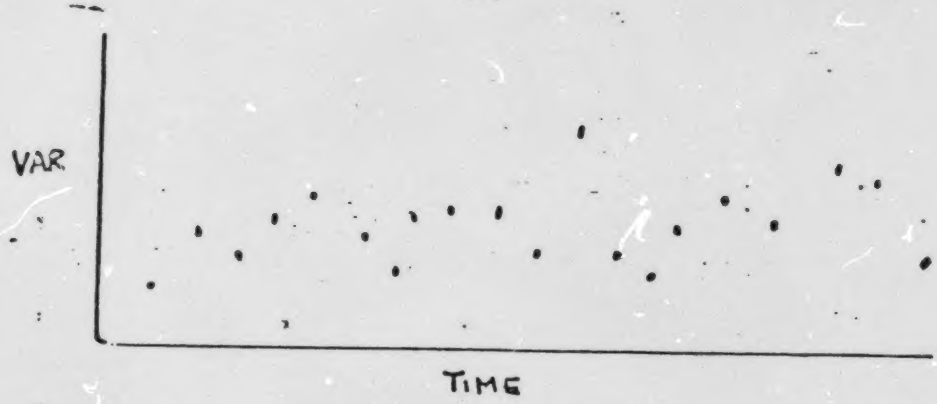


FIGURE 35-a Fig 3.33 a

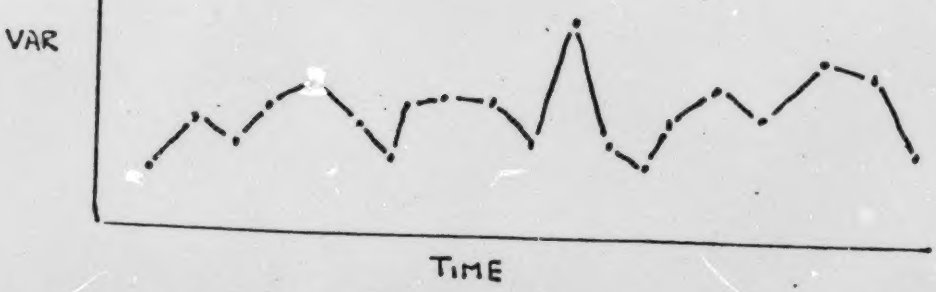
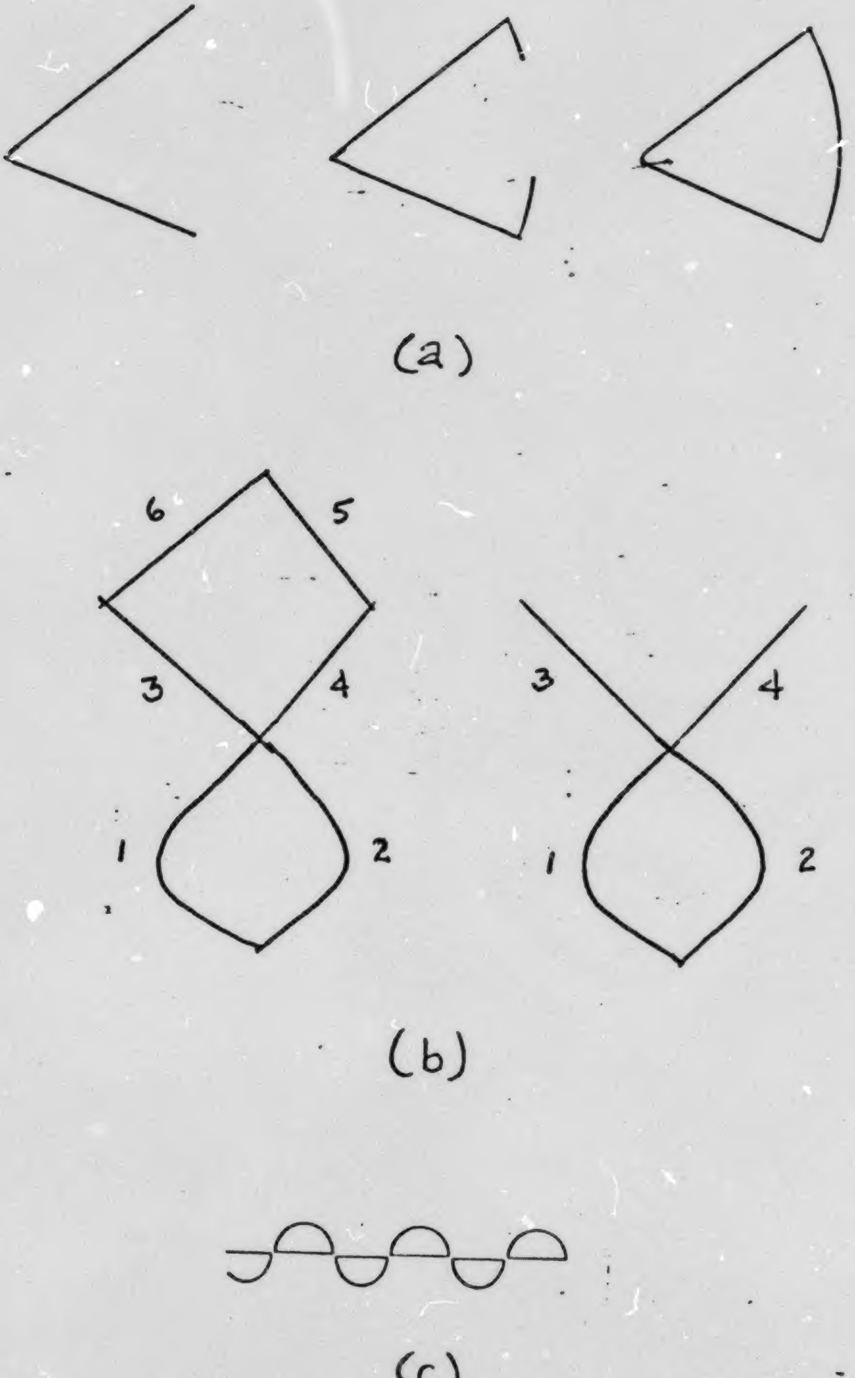


FIGURE 35-b Fig 3.33 b

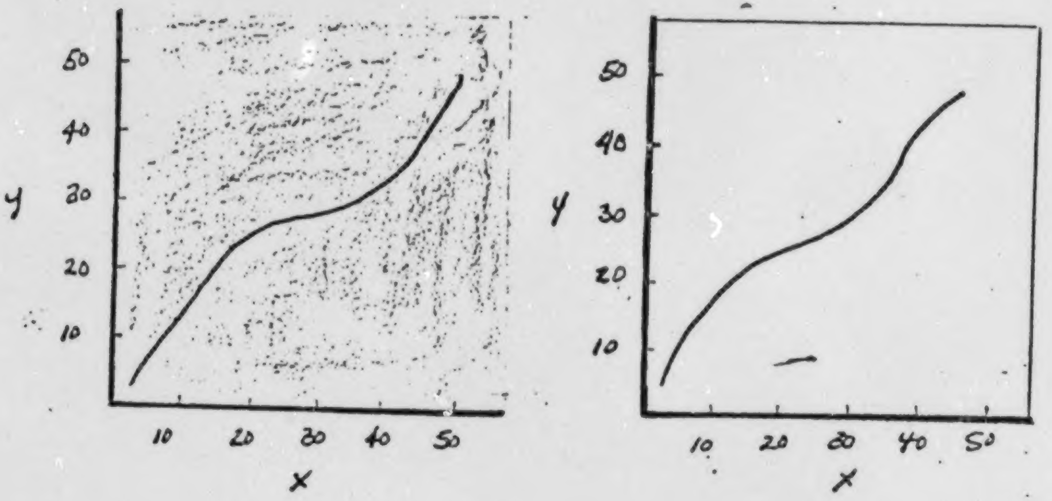
FIGURE ~~36~~ 35


(a)
6
5
3
4
1
2
3
4
1
2
(b)
(c)

388 Fig 3.34

3.3, 3.5

FIGURE 37



(a) (b)

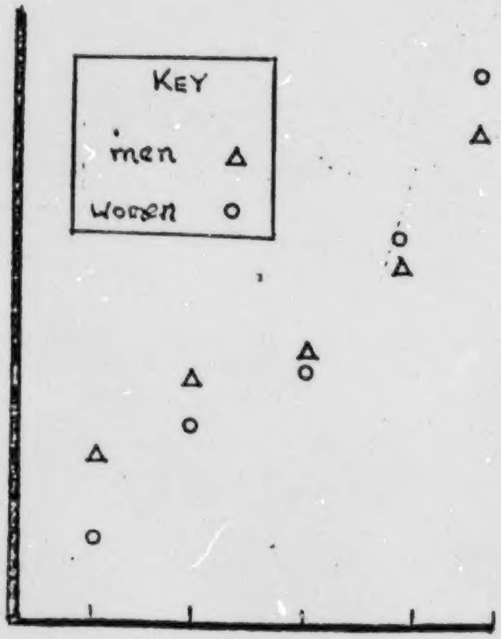


Fig 3.36

FIGURE 38

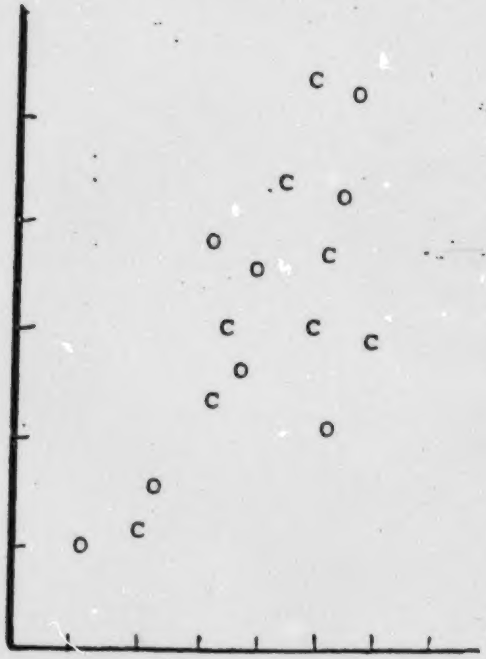
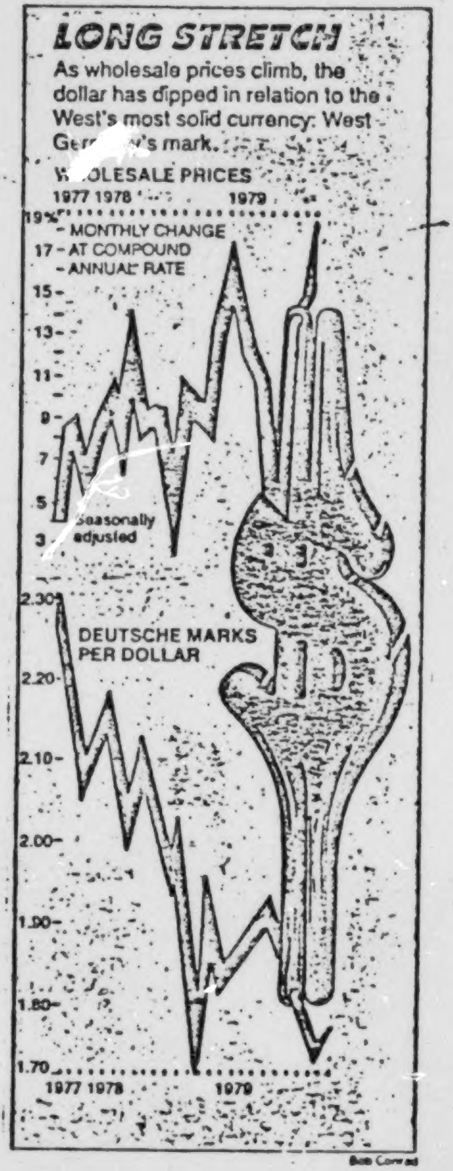


Fig 3.37

FIGURE 39



LONG STRETCH

As wholesale prices climb, the dollar has dipped in relation to the West's most solid currency: West Germany's mark.

Bob Conrad

FIGURE 40

Fig. 3, 38

Fig 3.39

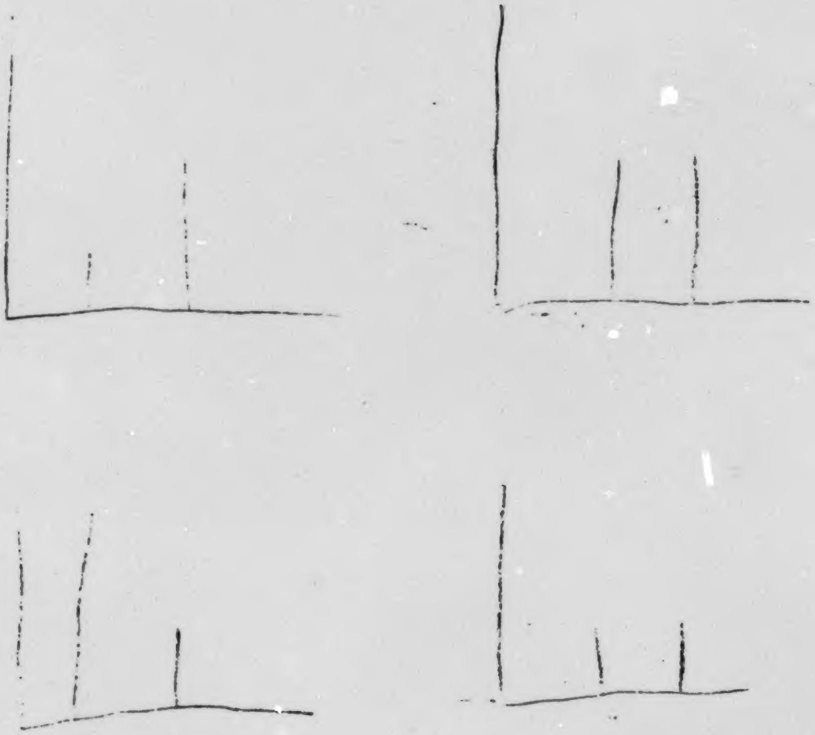


fig 3.39

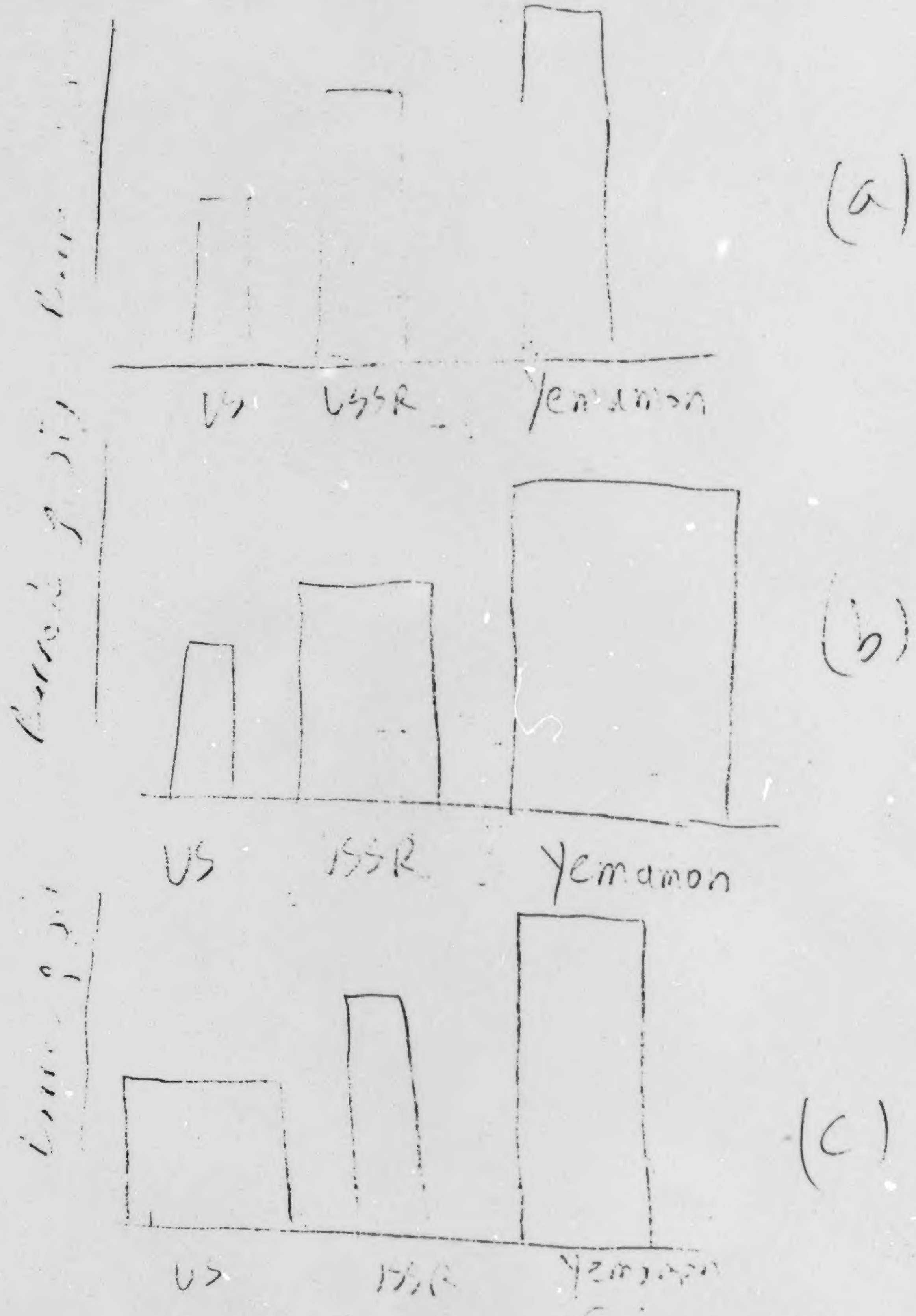


Fig 3.40

Table 3.14

Integral Dimensions

Object	Dimensions	Experimental Task	Reference
Munsell	Brightness	Free classification	Handle & Imai (1972)
Chip			Hyman & Well (1968)
	Saturation	Speeded classification	Garner & Felfoldy (1970)
Dot	Horizontal position	Speeded classification	Garner & Felfoldy (1970)
	Vertical position		
Ellipse	Eccentricity size	Absolute judgement	Egeth & Pachella (1969)
Rectangle	Length	Relative coding	Dykes & Cooper (1978)
	Width	Absolute judgement	Felfoldy (1974)
Obtuse	Height	Free classification	Somers & Pachella (1978)
Triangle	Length of right side		

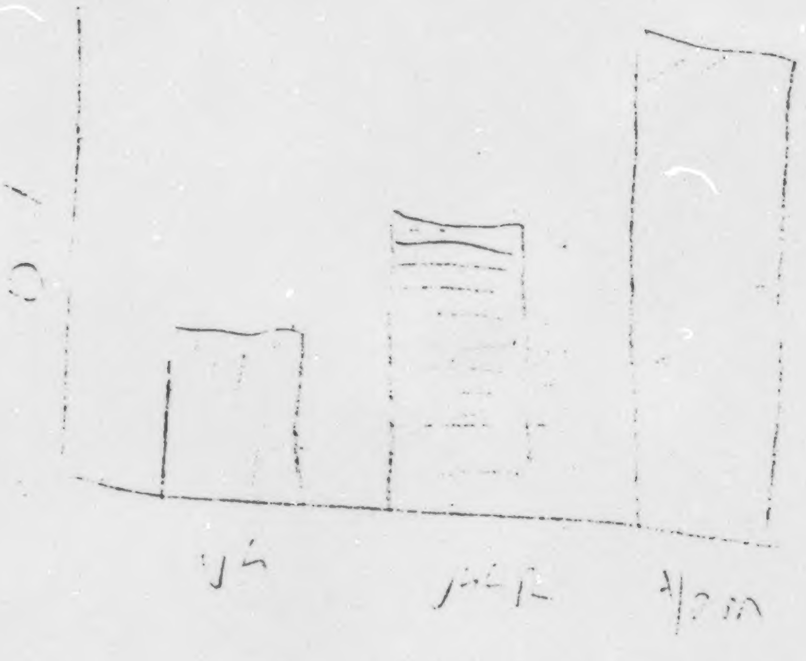
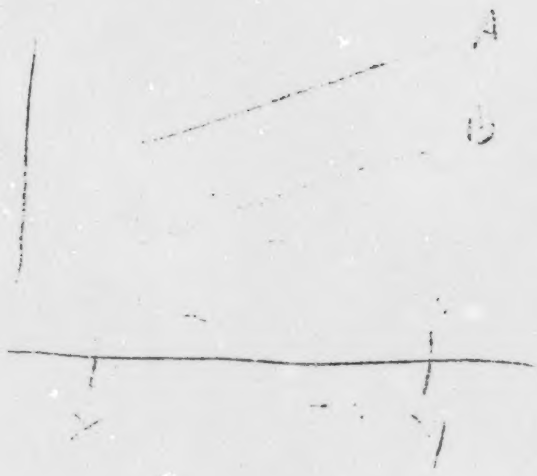
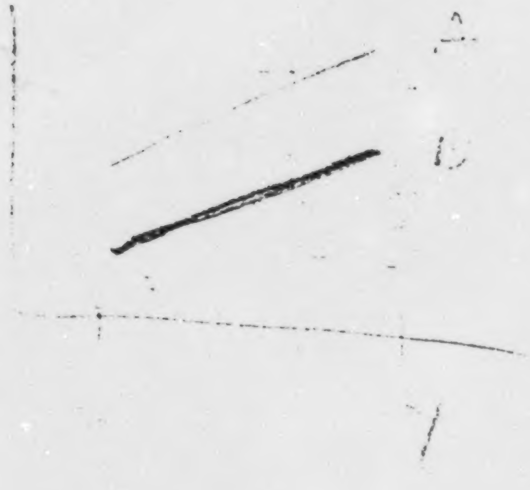


Fig 3.40(d)



(a)



(b)

3.41

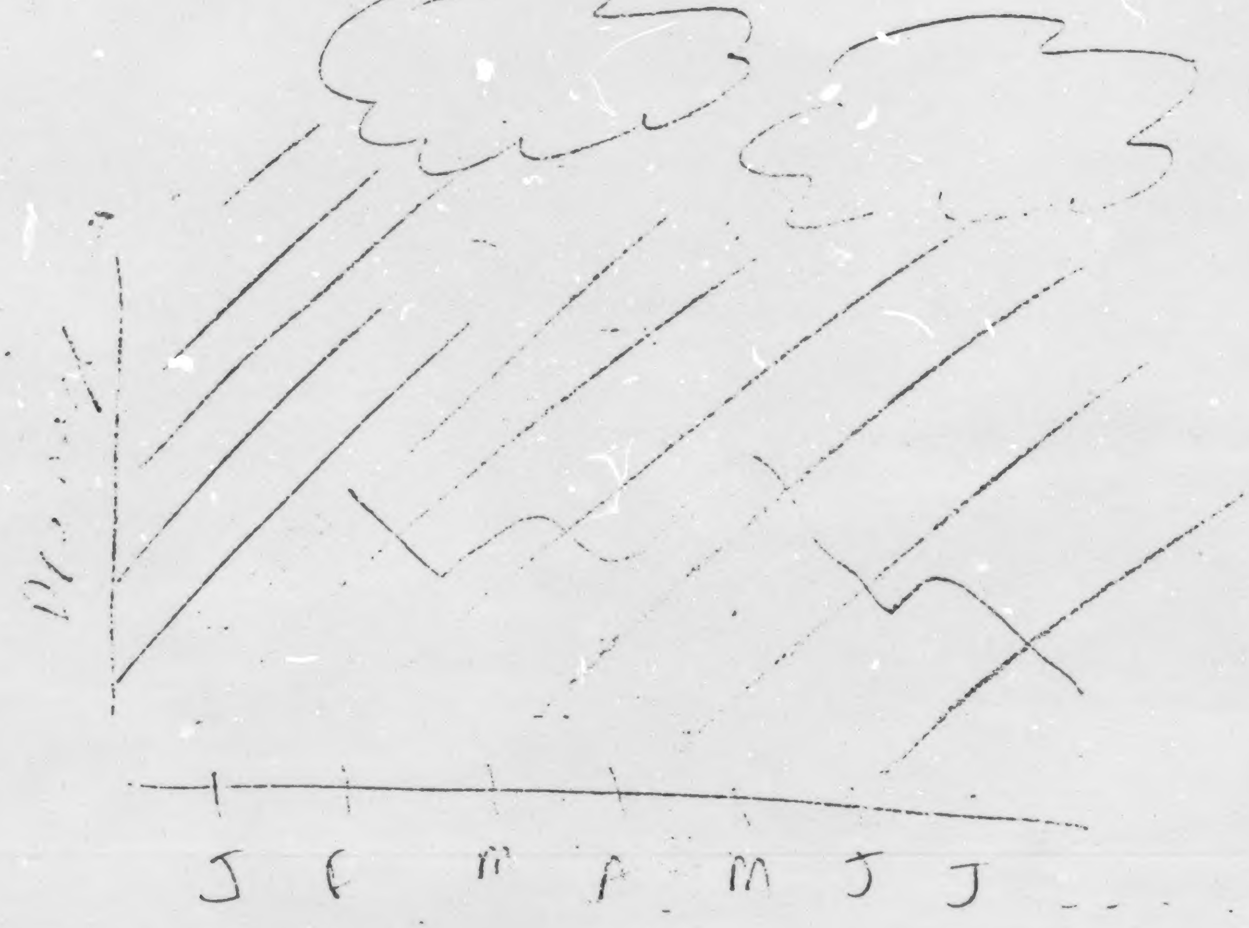


Fig 3.42
(a)

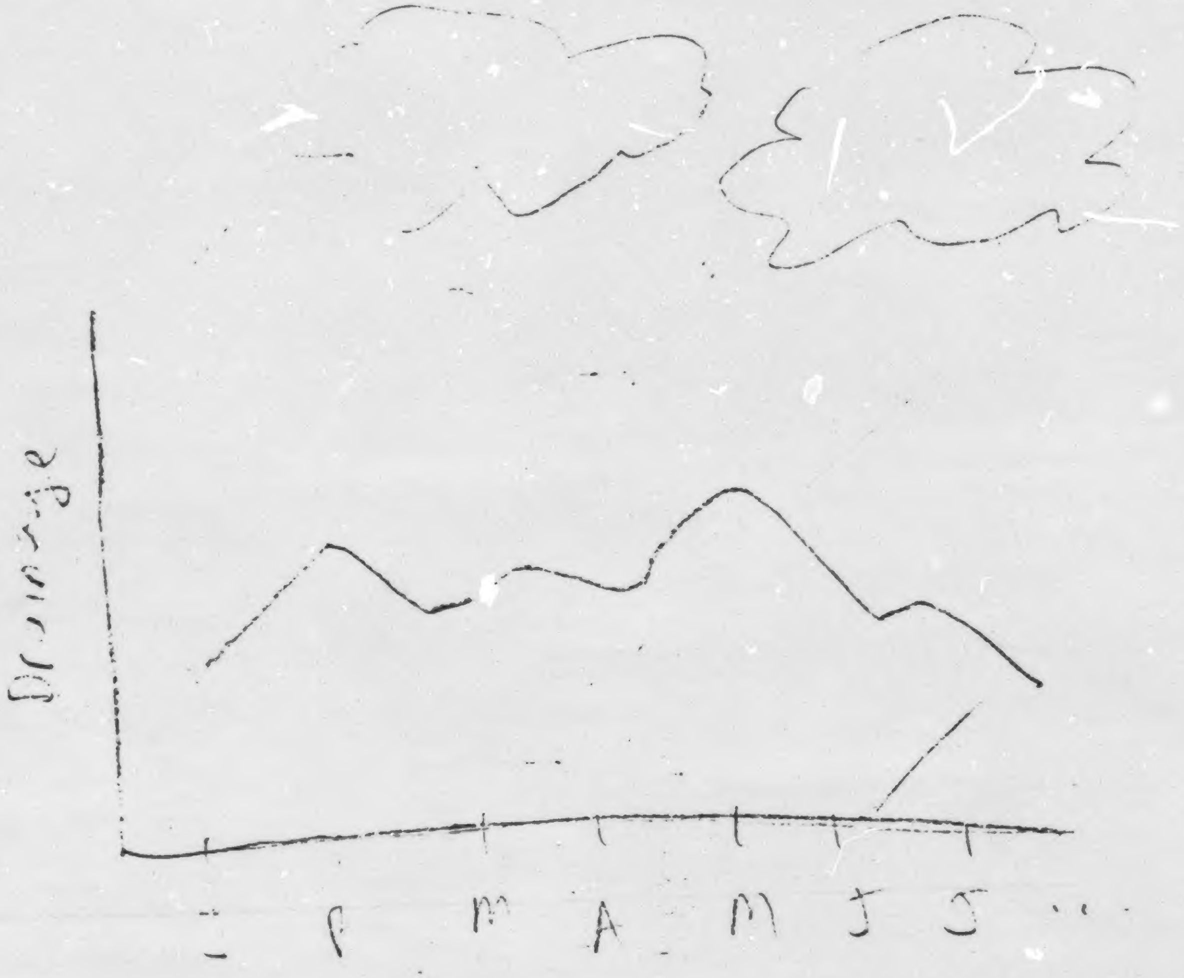


Fig 9.42 (b)

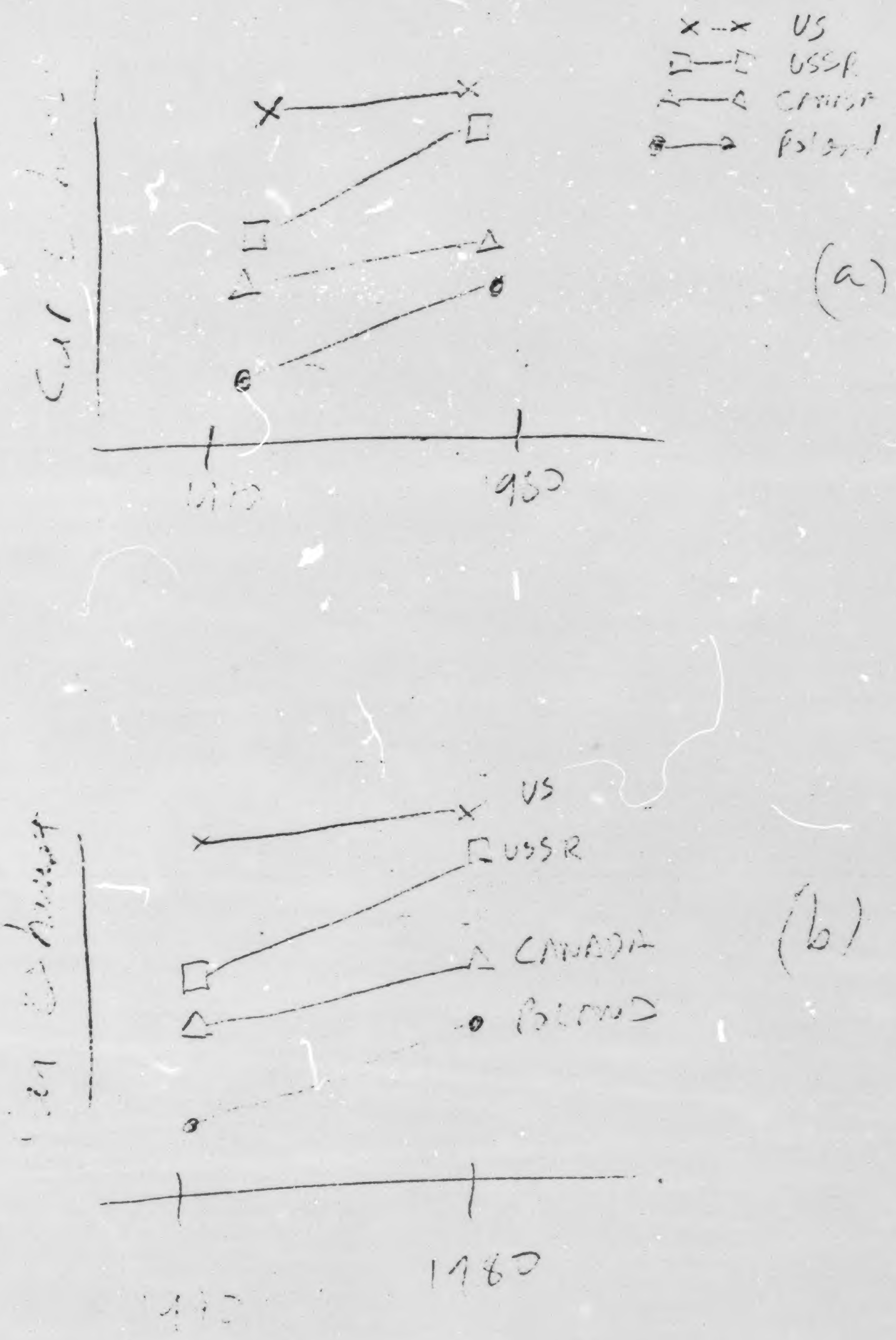


Fig 3.43

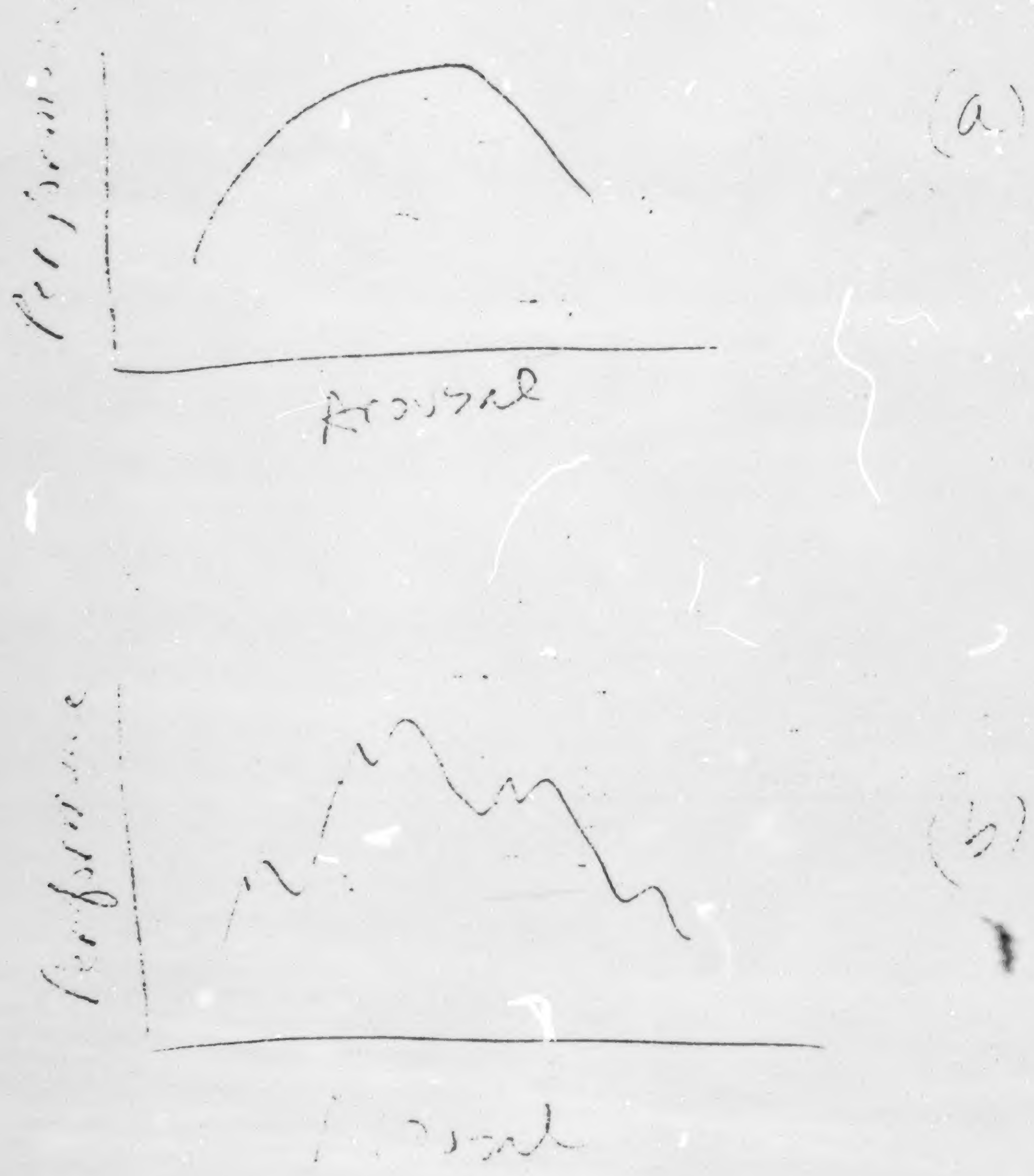
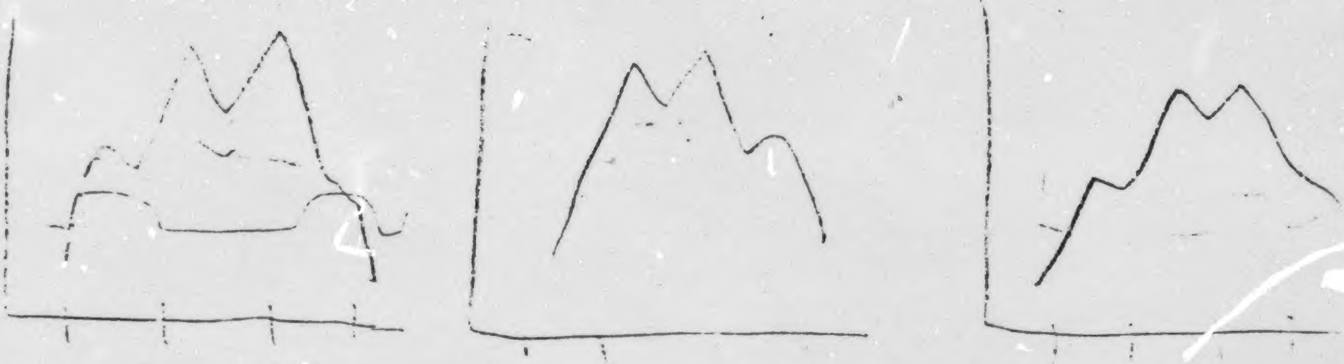


Fig 3.4

[illegible] 1493



(a) (b) (c)

Fig [illegible]

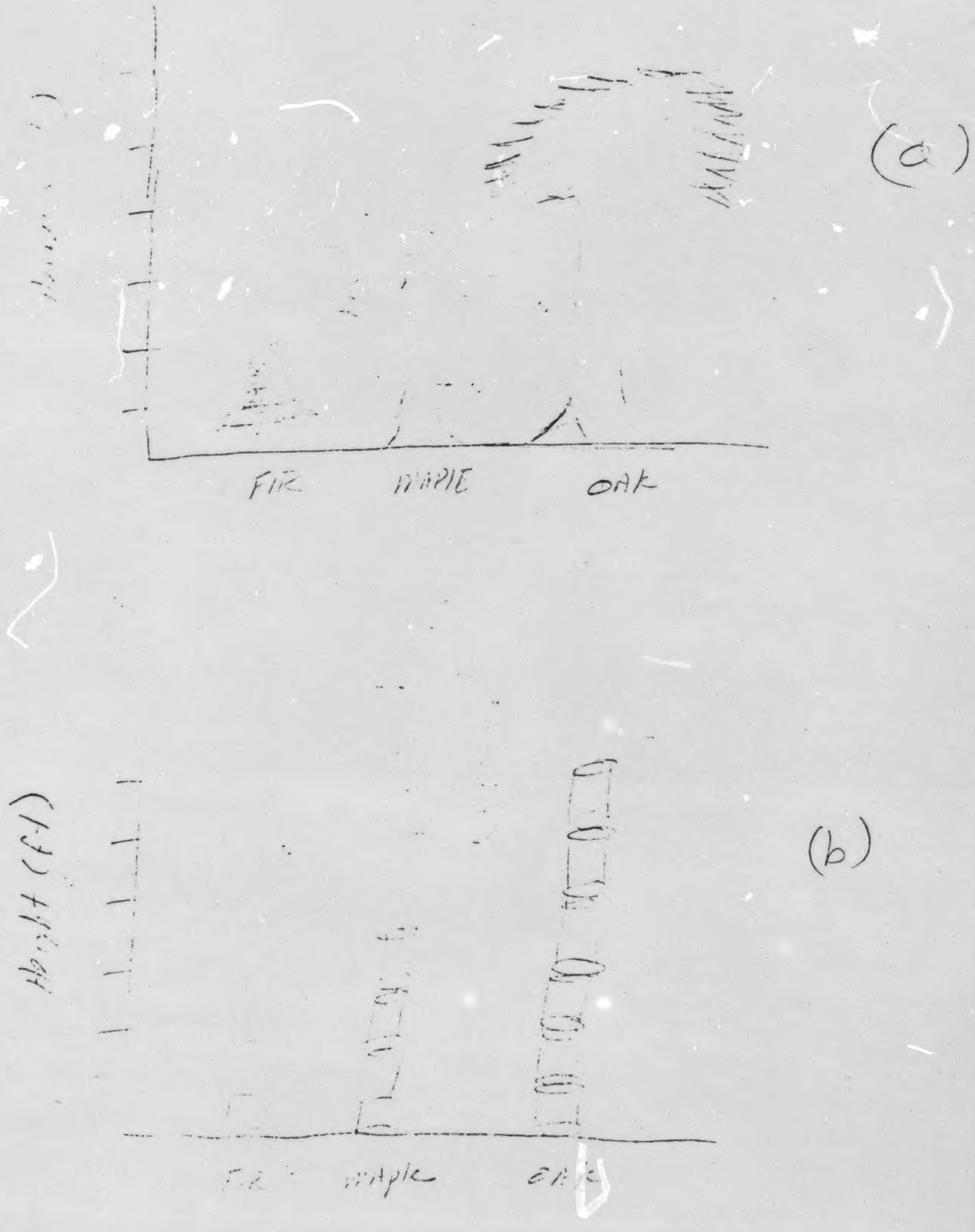

(a)
FIR
MAPLE
OAK
Height (ft)
(b)
Fir
maple
oak

Fig 4.2

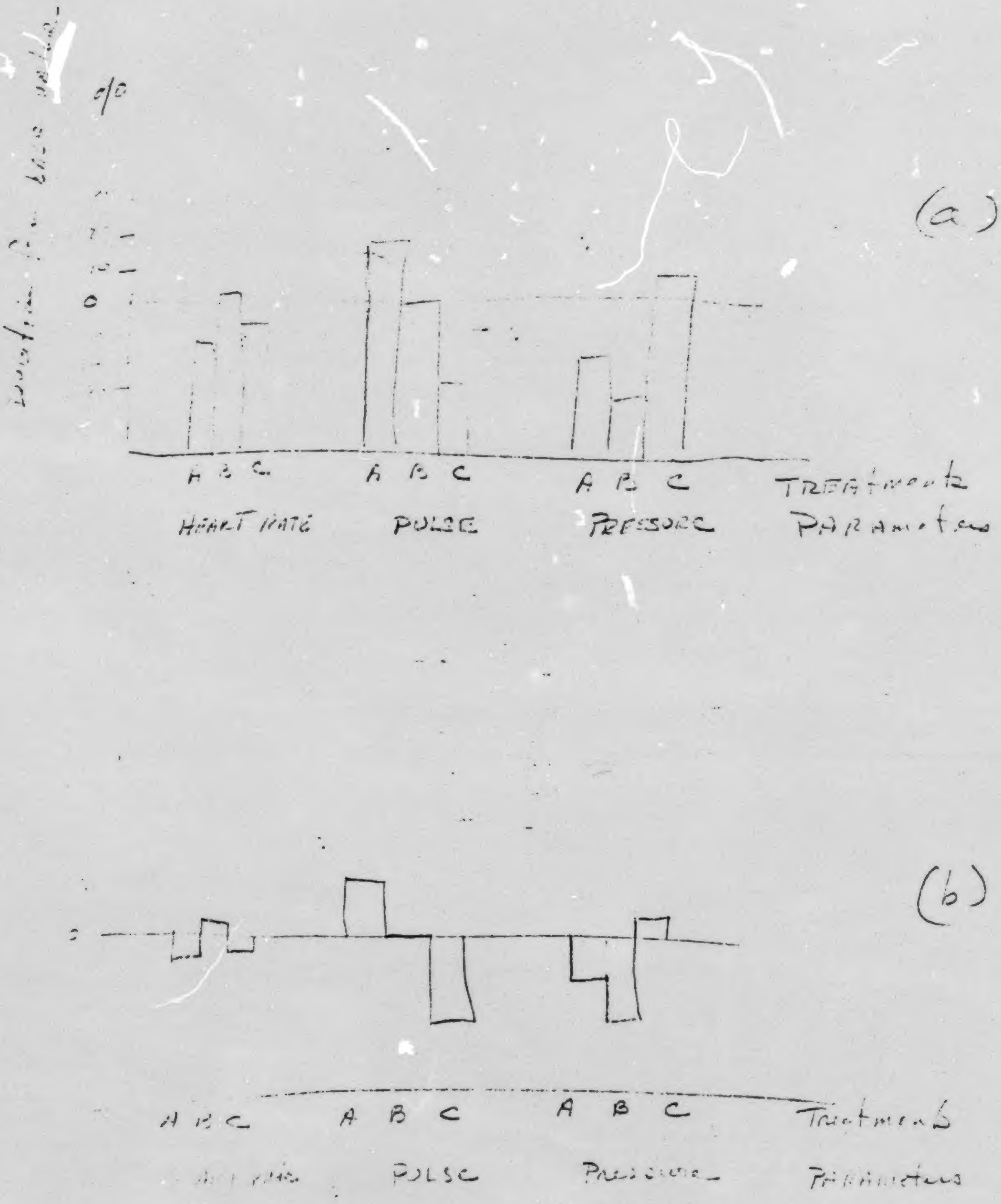

deviation from base value
%
0
(a)
A B C
A B C
A B C
Treatments
HEART RATE
PULSE
PRESSURE
Parameters
(b)
0
A B C
A B C
A B C
Treatments
Heart rate
PULSE
Pressure
Parameters

FIG 4.3

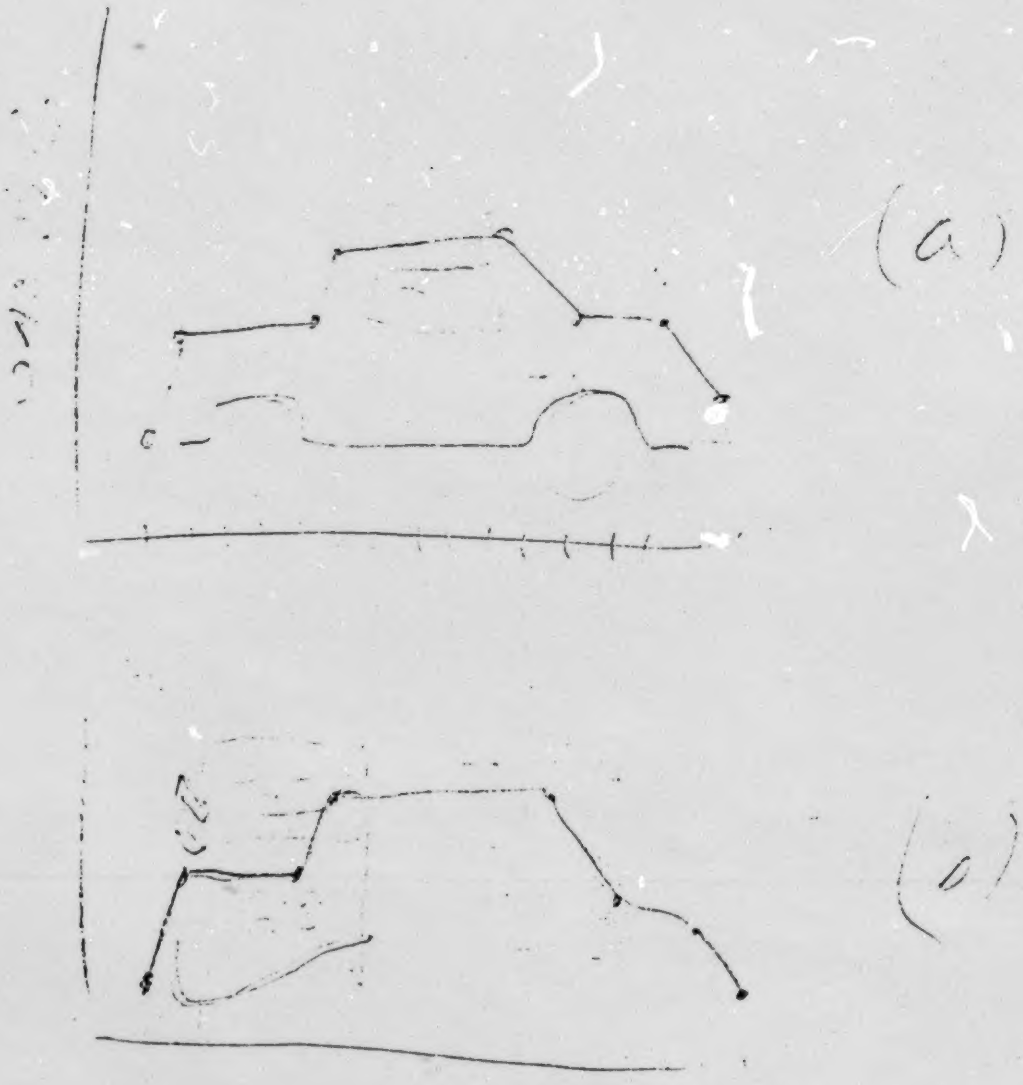
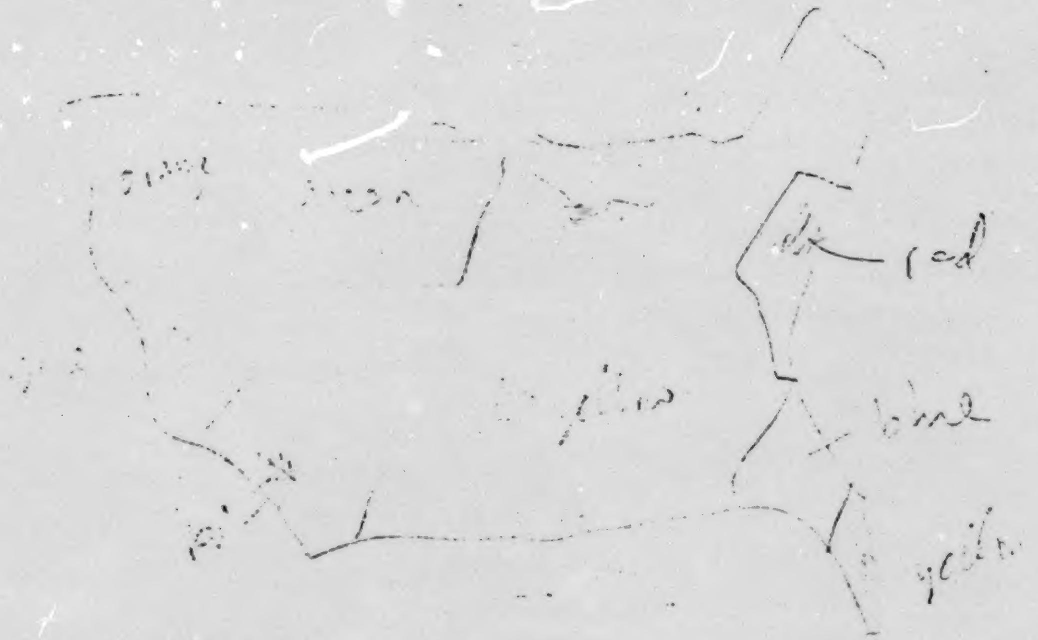


Fig 4.1


orange
green
red
plum
blue
yellow

red [illegible]
blue [illegible]
yellow [illegible]
orange [illegible]

Fig 4.4

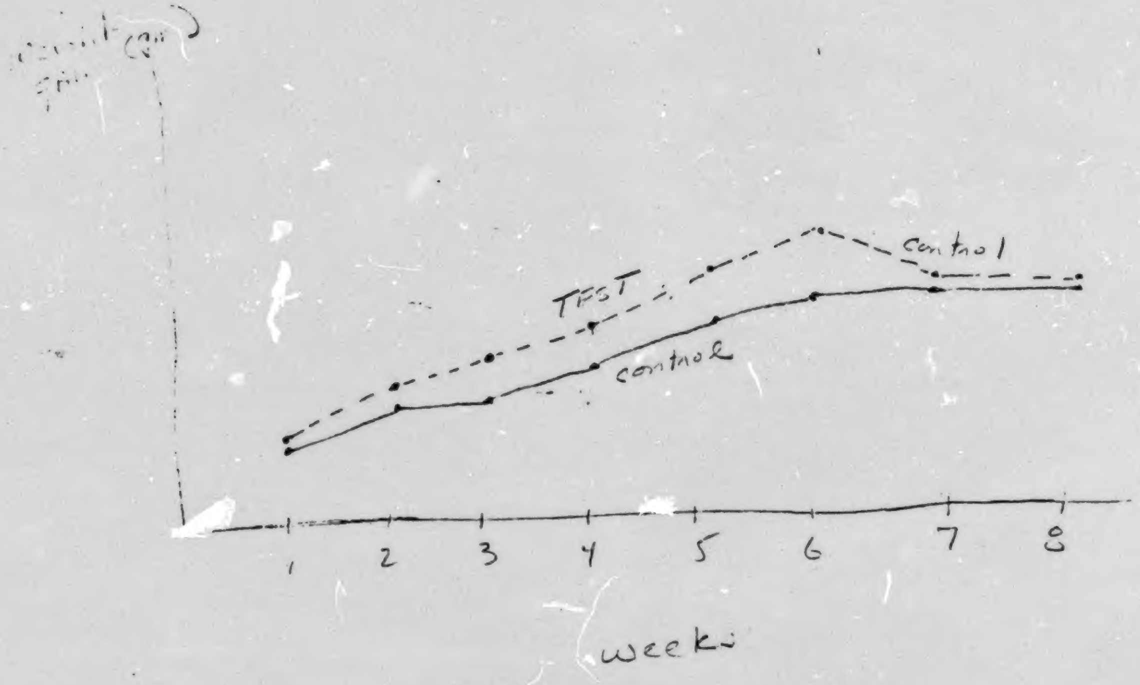


FIG 4.5

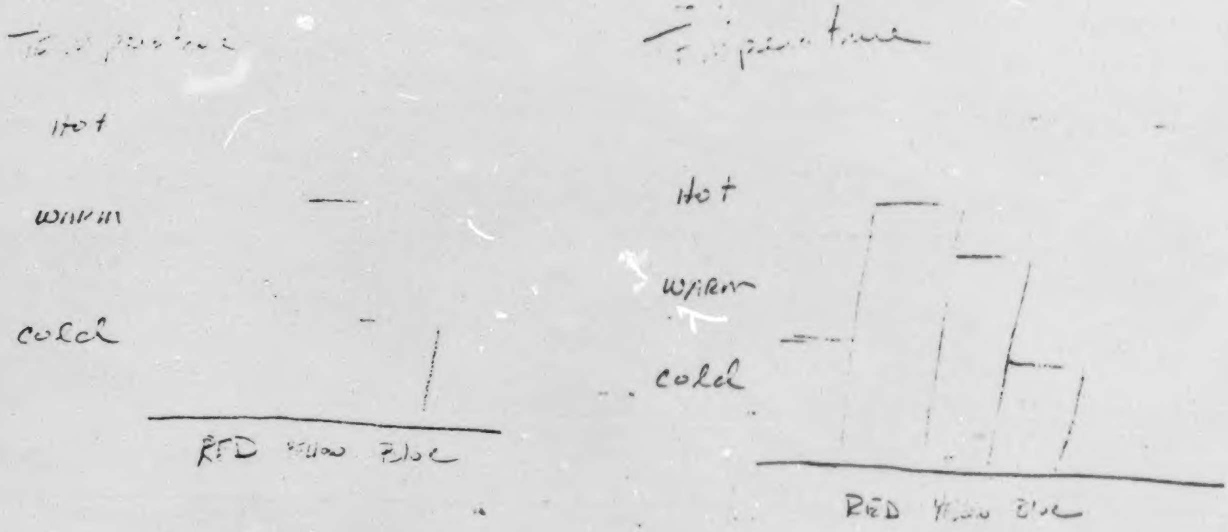

Temperature
hot
warm
cold
RED Yellow Blue
Temperature
Hot
warm
cold
RED Yellow Blue

FIG 4.6

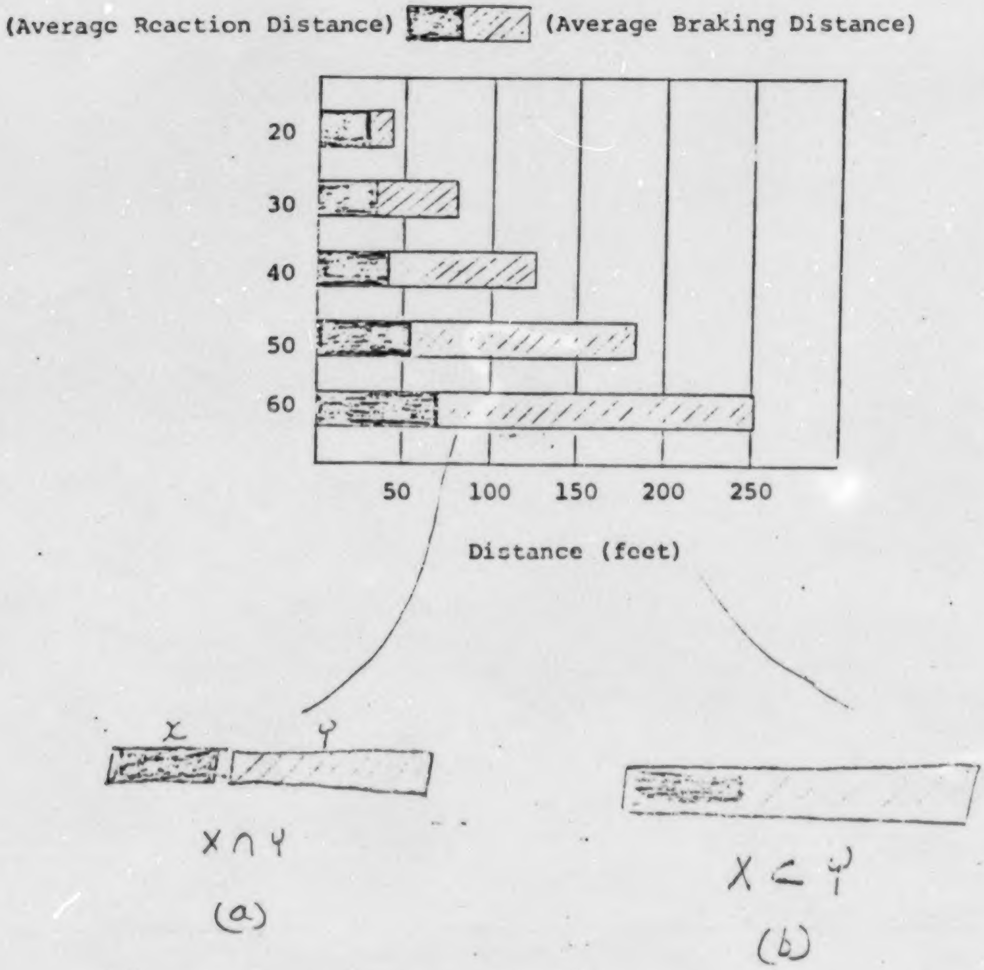
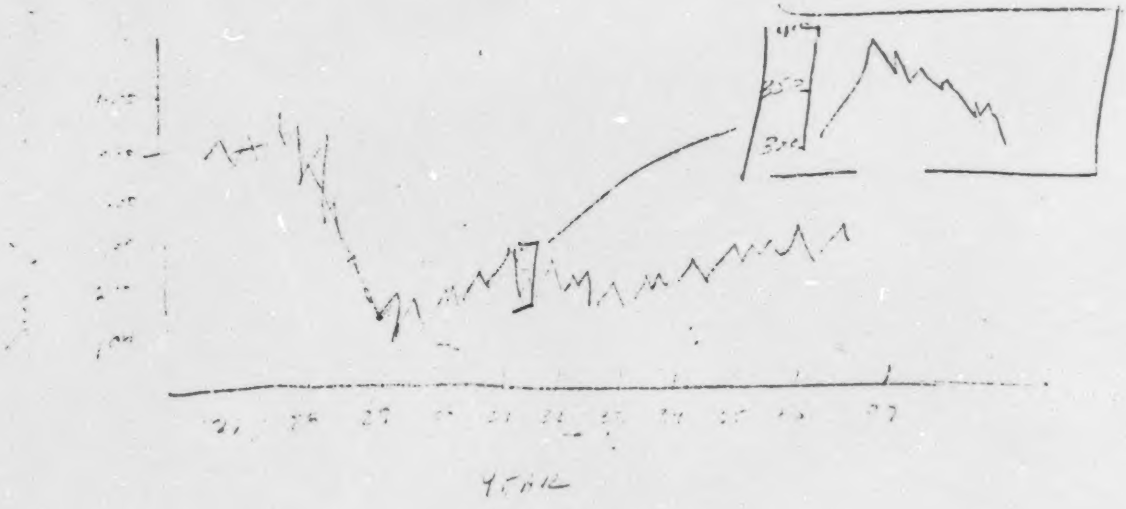
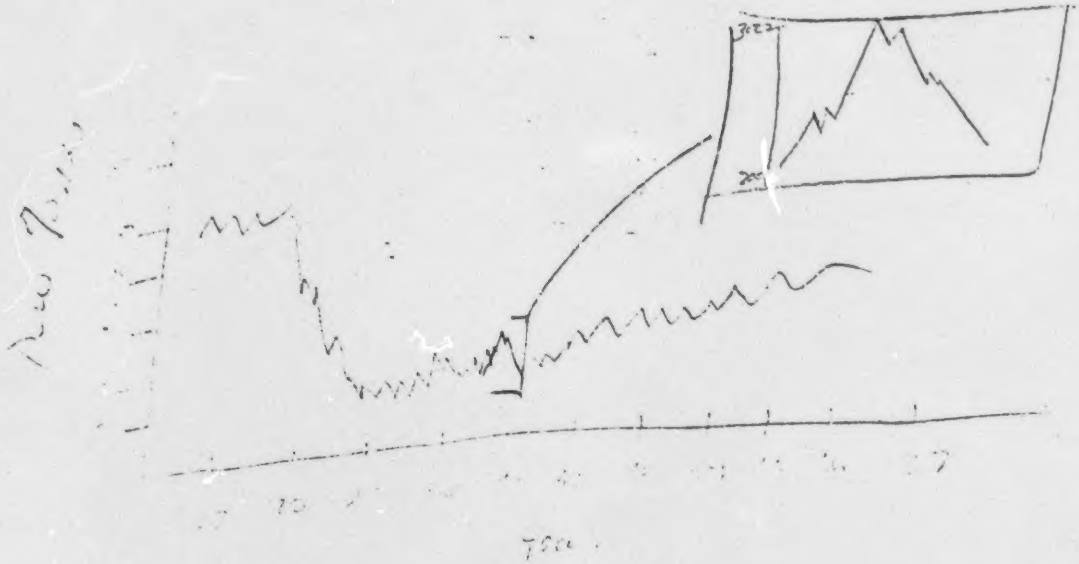

(Average Reaction Distance)
(Average Braking Distance)
20
30
40
50
60
50
100
150
200
250
Distance (feet)
x
y
X ∩ Y
(a)
X ⊂ Y
(b)

Fig 4.7

fig 4.5

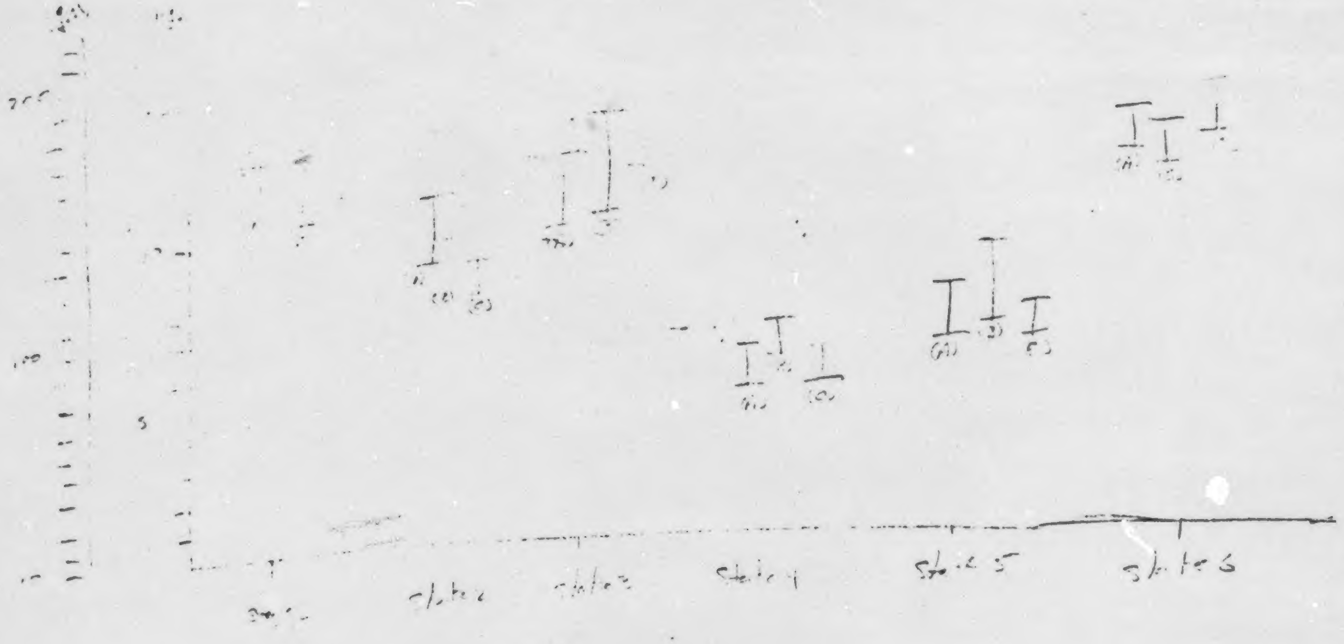


(a)

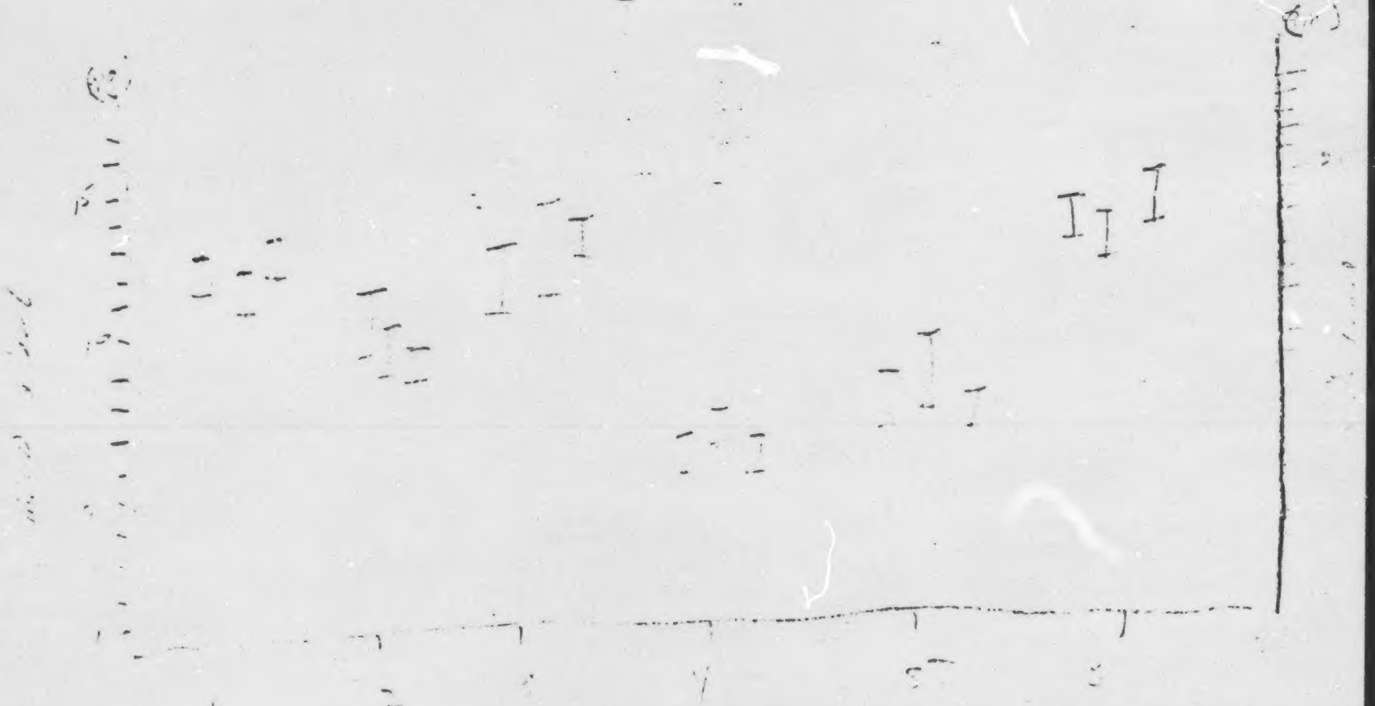


(b)

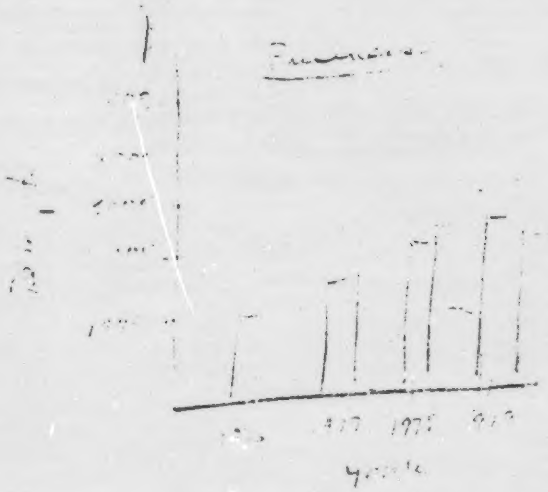
FIG 4.9



(a)



(b)



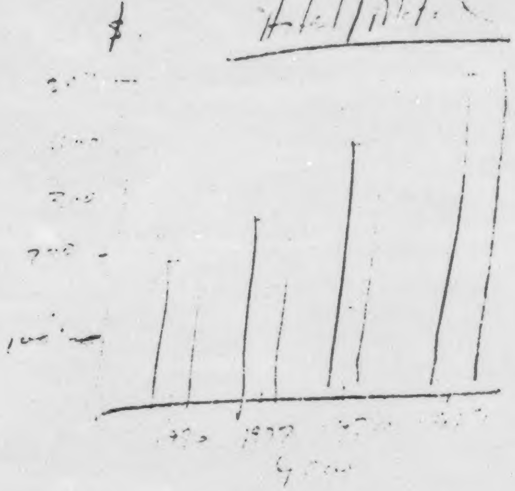


Fig 4.11

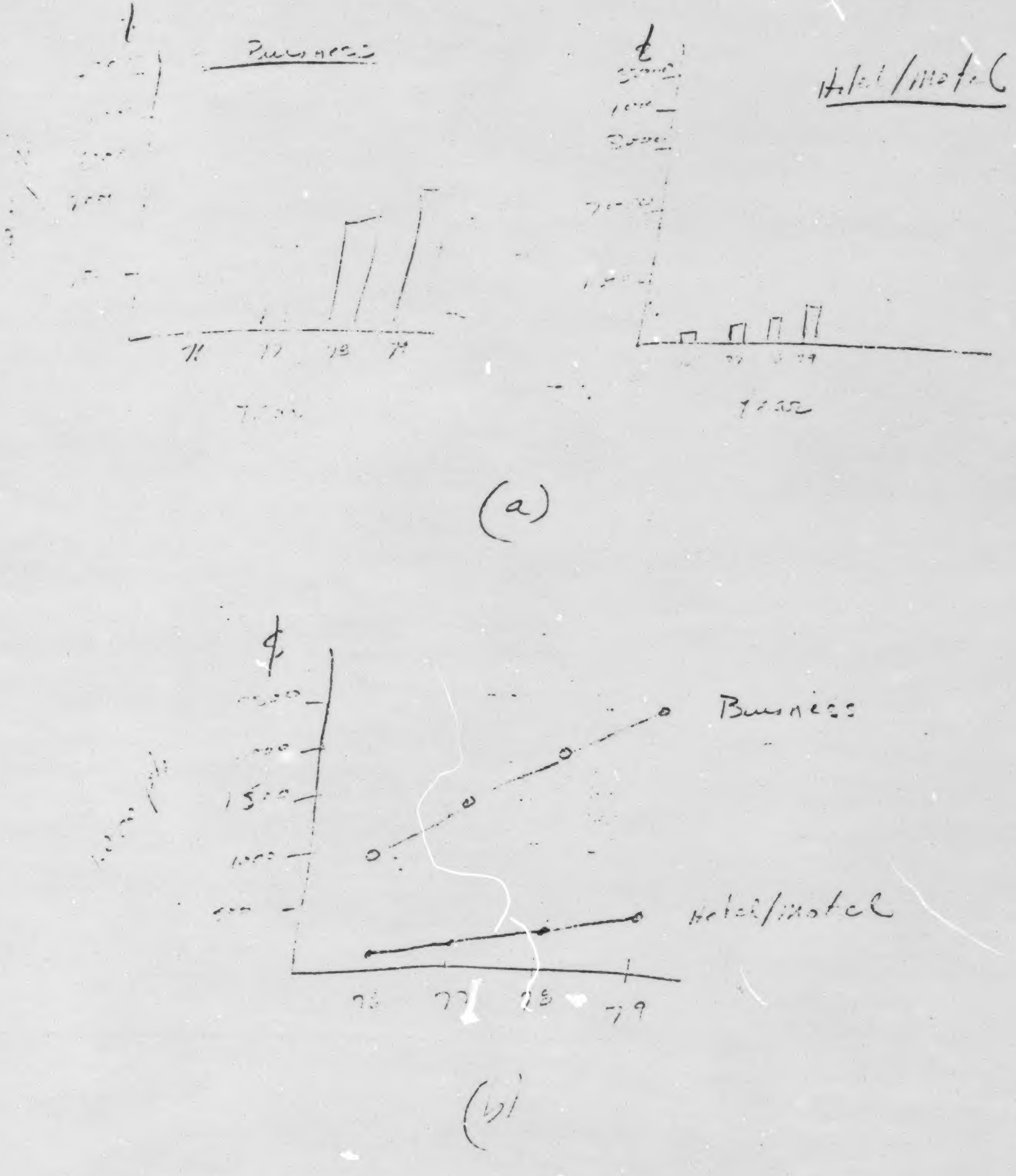

$
Business
$
Hotel/motel
76
77
78
79
Years
(a)
$
1500
1000
Business
Hotel/motel
76
77
78
79
(b)

Fig 4.12

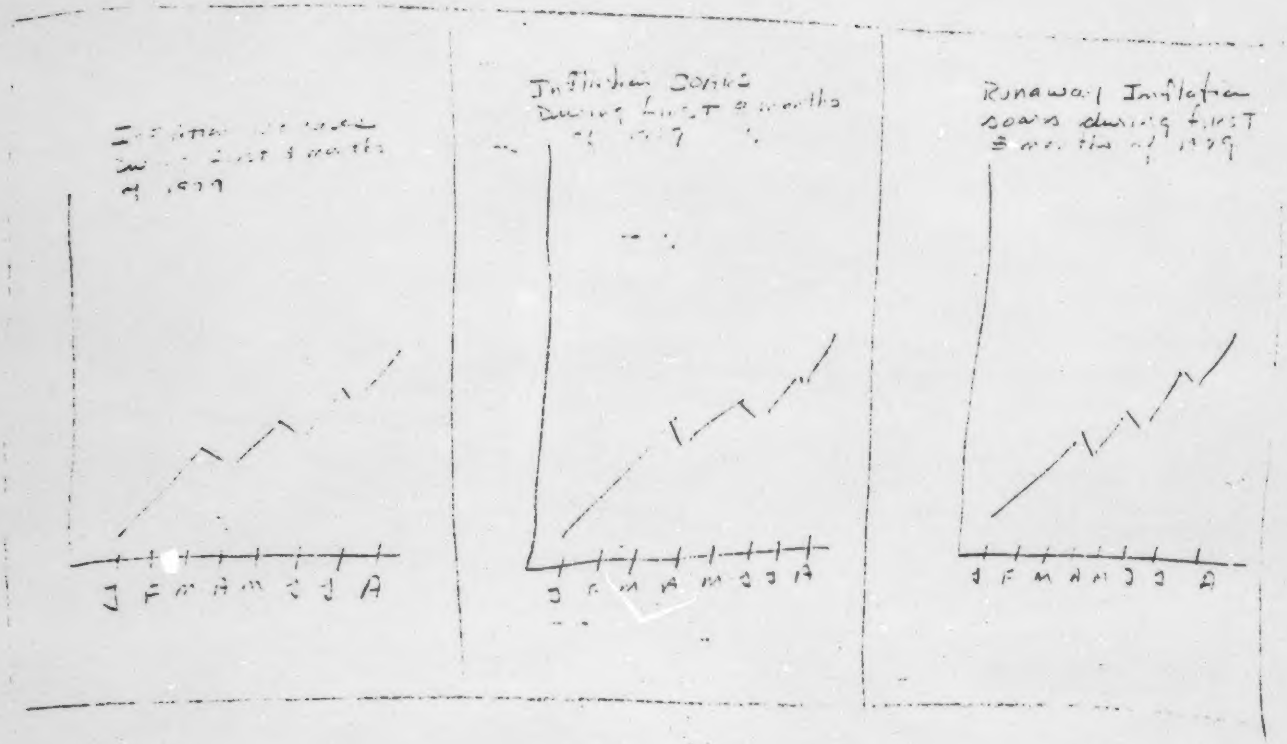
Inflation [illegible] during first 8 months of 1979
J F M A M J J A
Inflation soars During first 8 months of 1979
J F M A M J J A
Runaway Inflation soars during first 8 months of 1979
J F M A M J J A

Fig 4.13

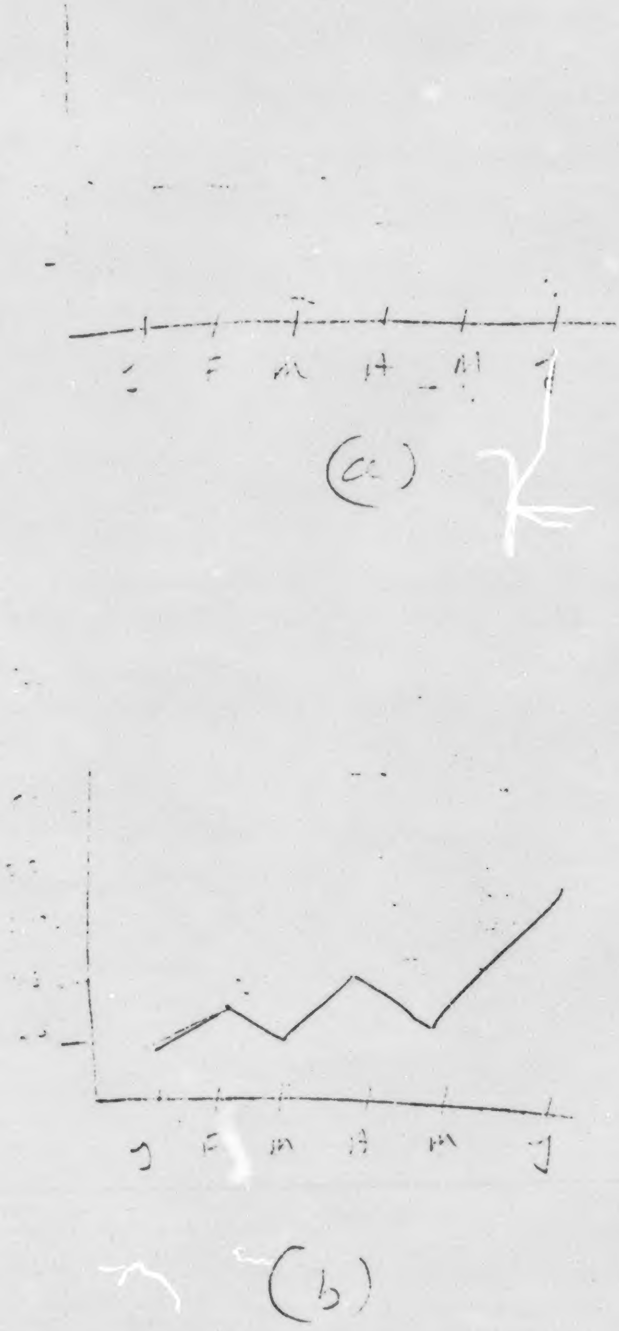
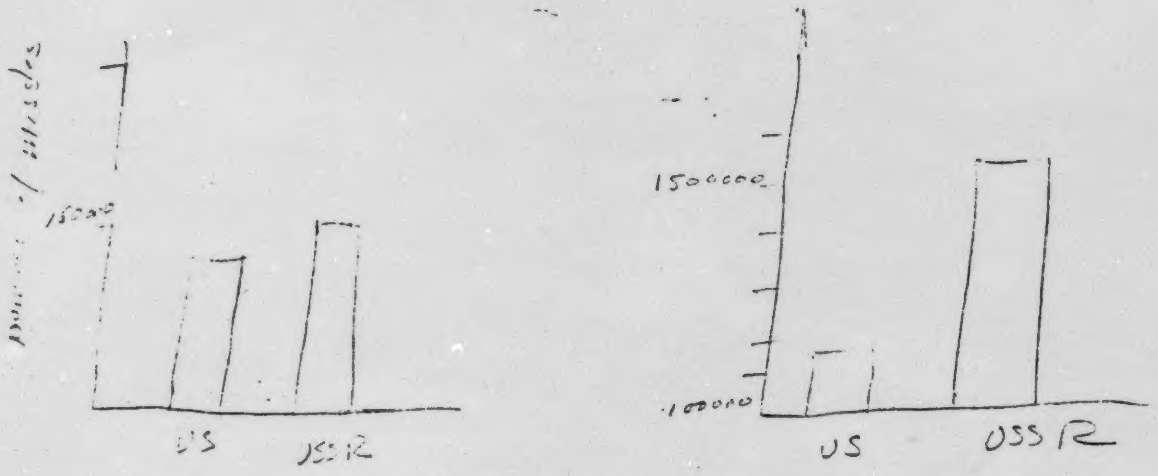


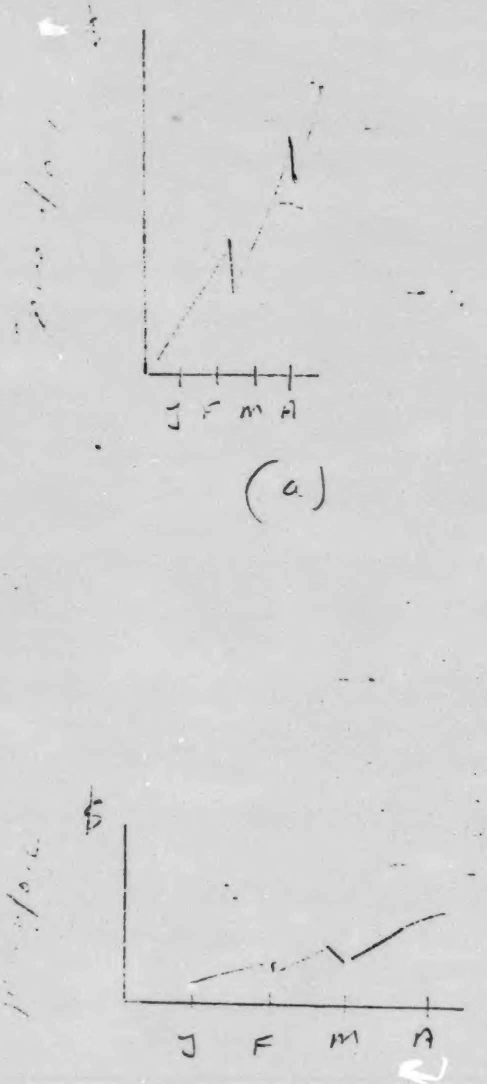
Fig 4. 14



(a)

(b)

FIG 4 15



(a)

(b)

FIG 4.16

To p1 - 28, ay 4

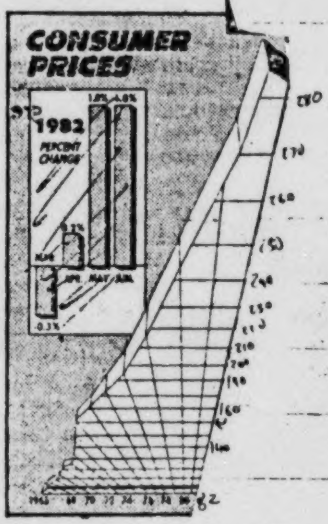
..ould be c..
..meals except those
travel.

a
r
it

x
d
d
of

es
ar
ig
ie

as
w
31.
nt
22



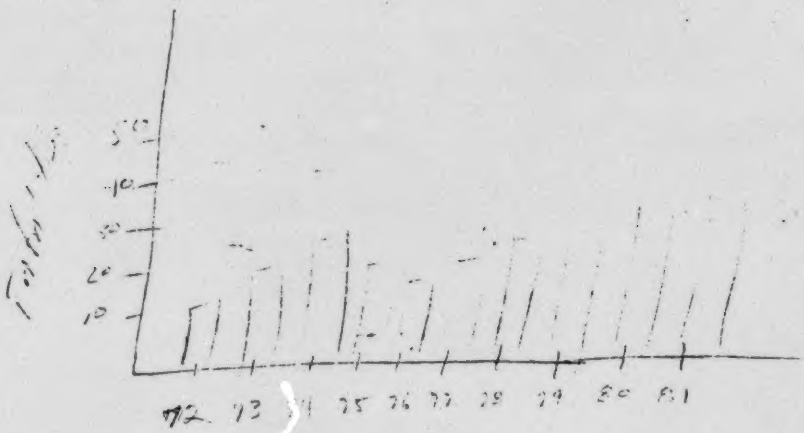
the monthly CPI
on a number tha

CPI:
1967=100

7
Figure 4.18

417

Fig 4.17

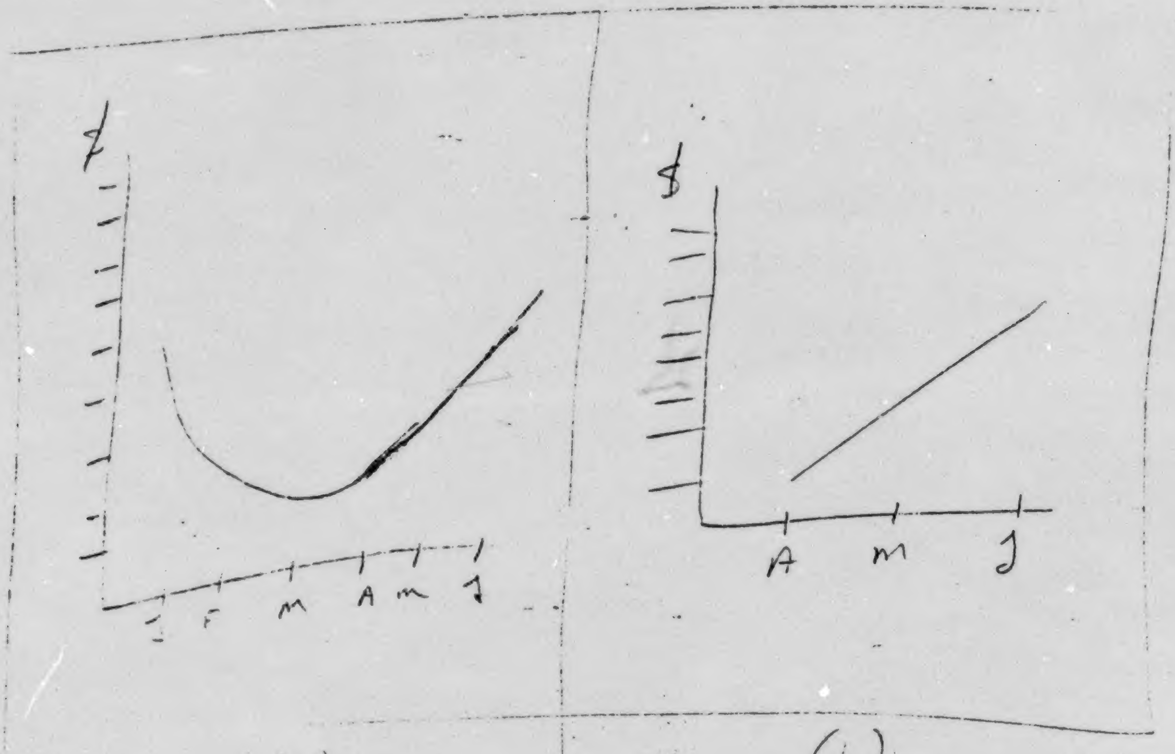


(a)

will draw!

(b)

Fig 4.18



(a) (b)

FIG 4.19

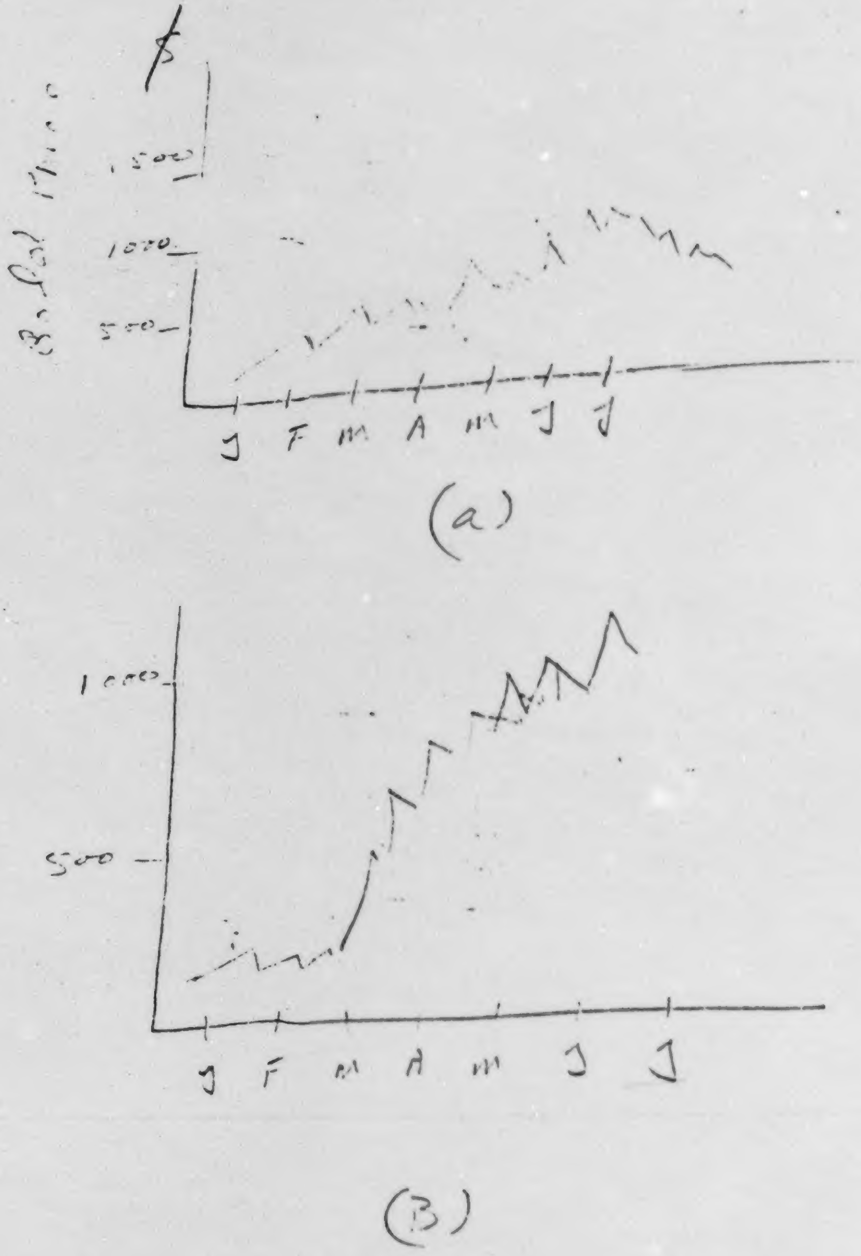


Fig 4.20

Table 5.1

Distribution of Questions for Each of the Different Operating Principles

Syntax

	Adequate Discriminability	Dimensional Structure	Gestalt Organization	Perceptual Distortion	Processing Limitations	Processing Priorities
No. of questions	11	1	9	1	6	11

Formal and Semantic

	Horizontal Mapping	Vertical Mapping	Schema Availability	Surface Compatibility
No. of questions	9	17	4	7

Pragmatics

	Contextual Compatibility	Invited Inference
No. of questions	4	5

Table 5.2

Distribution of Questions for Each of the Graphic Constituents and Their Combinations

Constituents

	Framework	Specifier	Labels	Background
No. of questions	17	19	28	4

Combinations of Constituents

	Frame - Specifiers	Frame - Specifiers - Labels
No. of questions	3	4

Table 5.3

Possible Outcomes of an Analysis of a Graph by Two Analysts

		Analyst 2	
		Problem	No Problem
Analyst 1	Problem	a	b
	No Problem	c	d

Table 5.4

Results From Analysis of Ten Graphs by Two Analysts

		Analyst 2	
		Problem	No Problem
Analyst 1	Problem	58	9
	No Problem	18	705

Table 5.5

Distribution of Graphs as a Function of the Sampling Scheme Categories

Content Area

	Math	Physical Science	Life Science	Social Science	Business	General Interest
No. of Graphs Analyzed	10	11	16	15	13	10

Audience

	Adult	Secondary	Primary	General
No. of Graphs Analyzed	40	18	7	10

Publication Format

	Journal	Textbook	General Reading	Newspaper	Magazine
No. of Graphs Analyzed	15	37	16	3	4

Visual Format

	Bar	Line	Pie	Other
No. of Graphs Analyzed	22	24	10	19

Table 5.6

Proportion of Faults for the Different Sampling Scheme Categories

Content Area

	Math	Physical Science	Life Science	Social Science	Business	General Interest
Faults/ Graph	1.2	1.9	1.2	1.7	2.8	1.4
	(10)*	(11)	(16)	(15)	(13)	(10)

Audience

	Adult	Secondary	Primary	General
Faults/ Graph	1.9	1.4	1.4	1.5
	(40)	(18)	(7)	(10)

Publication Format

	General	Journal	Magazine	Newspaper	Textbook
Faults/ Graph	2.1	1.4	1.25	0.33	1.81
	(16)	(15)	(4)	(3)	(37)

Visual Format

	Bar	Line	Pie	Other
Faults/ Graph	1.7	1.5	1.2	2.1
	(22)	(24)	(10)	(19)

*Number of graphs in parentheses

Table 5.7

Distribution of Faults Per Question Set as a Function of the Different Levels of Analysis and Operating Principles

(a)
Levels of Analysis

	Syntax	Semantics	Pragmatics	Formal
Proportion of Faults	1.5 (39)*	1.2 (11)	1.0 (9)	1.8 (26)

*Number of questions in a set are shown in the parentheses.

(b)
Operating Principles

Syntax

	Adequate Discriminability	Dimensional Structure	Gestalt Organization	Perceptual Distortion	Processing Limitations	Processing Priorities
Proportion of Faults	1.6 (11)	0 (1)	1.9 (9)	2.0 (1)	1.3 (6)	1.3 (11)

Formal and Semantic

	Internal Mapping	External Mapping	Schema Availability	Surface Compatibility
Proportion of Faults	1.0 (9)	2.2 (17)	1.2 (4)	1.1 (7)

Pragmatics

	Contextual Compatibility	Invited Inference
Proportion of Faults	0.9 (4)	0.6 (5)

Table 5.8

Distribution of Fault Proportions as a Function of the Different Graphic Constituents

	Constituent						
	Background	Label	Frame	Specifier	Fra-Spec	LA-FR-Spec	Mult-Fra
Proportion of Faults	1.0 (4)*	1.7 (28)	1.0 (17)	2.3 (19)	2.0 (3)	1.7 (4)	0.3 (10)

*Number of questions in a set is shown in parentheses.

Table 5.9

Breakdown of Fault Proportion for Specifier and Frame-Specifier Combination in Terms of the Different Operating Principles

Operating Principle

	Internal Mapping	External Mapping	Surface Compatibility	Adequate Discriminability	Gestalt Organization	Perceptual Distortion	Processing Limitations	Processing Priorities
Proportion of Faults	1.00	0	0	0	0	0	0	0
	0.02	0.37	0.09	0.23	0.02	0.05	0.14	0.07

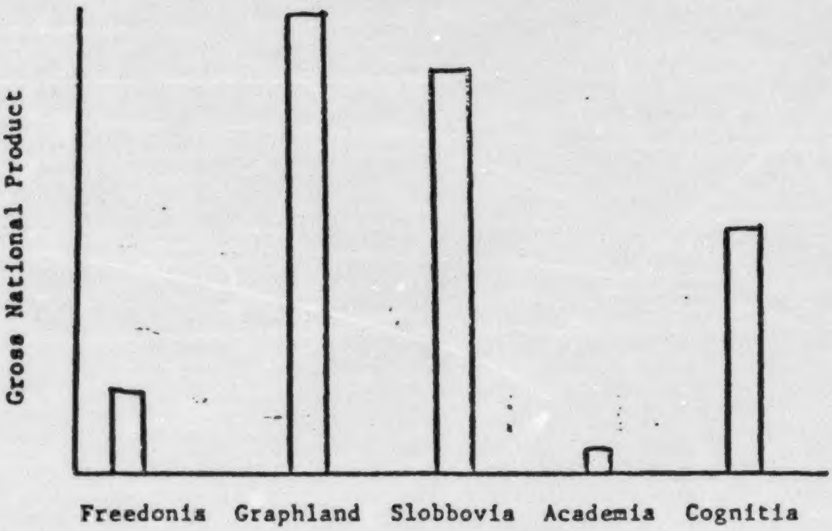


Figure 1.

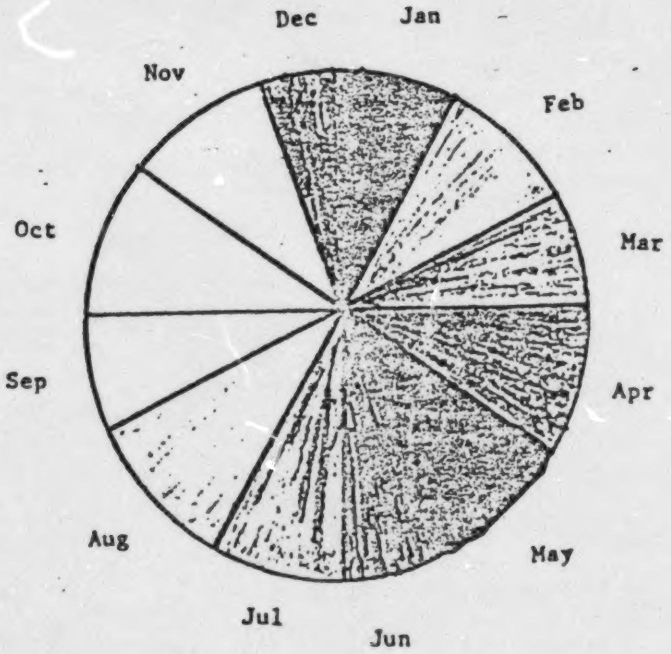


Figure 2.

FIGURE 1

Distribution of Total Sample of Charts and Graphs Visual Format

Field	Audience	Format	Bar 1	Line 2	Pie 3	Other 4
Mathematics	Adult	Journal (MAJ)	1	4		2
		Textbook (MAT)		2		1
		General Reading (MAG)				
	Secondary	Textbook (MST)	4	4	2	2
		General Reading (MSG)				
	Pre-Secondary	Textbook (MPT)				1
		General Reading (MPG)				
Physical Sciences	Adult	Journal (PAJ)	2	4		
		Textbook (PAT)		14		
		General Reading (PAG)	2	1		
	Secondary	Textbook (PST)	3	4	1	4
		General Reading (PSG)				
	Pre-Secondary	Textbook (PPT)	1	1		
		General Reading (PPG)				
Life Sciences	Adult	Journal (LAJ)	13	15		4
		Textbook (LAT)	2	6	1	3
		General Reading (LAG)	5	8	2	5
	Secondary	Textbook (LST)	2	1		1
		General Reading (LSG)				
	Pre-Secondary	Textbook (LPT)		1		1
		General Reading (LPG)				
Social Sciences	Adult	Journal (SAJ)	5	7		1
		Textbook (SAT)	2	3		2
		General Reading (SAG)	2	2		1
	Secondary	Textbook (SST)	5	3	3	2
		General Reading (SSG)				
	Pre-Secondary	Textbook (SPT)	2		1	
		General Reading (SPG)				
Business	Adult	Journal (BAJ)	8	11	1	5
		Textbook (BAT)	1	2		
		General Reading (BAG)	6	3	2	3
	Secondary	Textbook (BST)	1	5	2	
		General Reading (BSG)				
	Pre-Secondary	Textbook (BPT)				
		General Reading (BPG)				
General Interest	General	Newpaper (GIN)	6	10		4
		Magazine (GIM)	13	21	2	8
		General Reading (GIG)	6	11		1

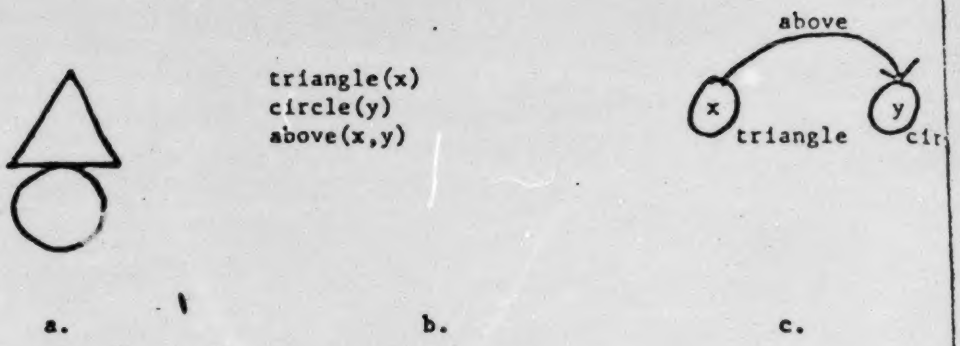

triangle(x)
circle(y)
above(x,y)
above
x
y
triangle
cir
a.
b.
c.

Figure 3.

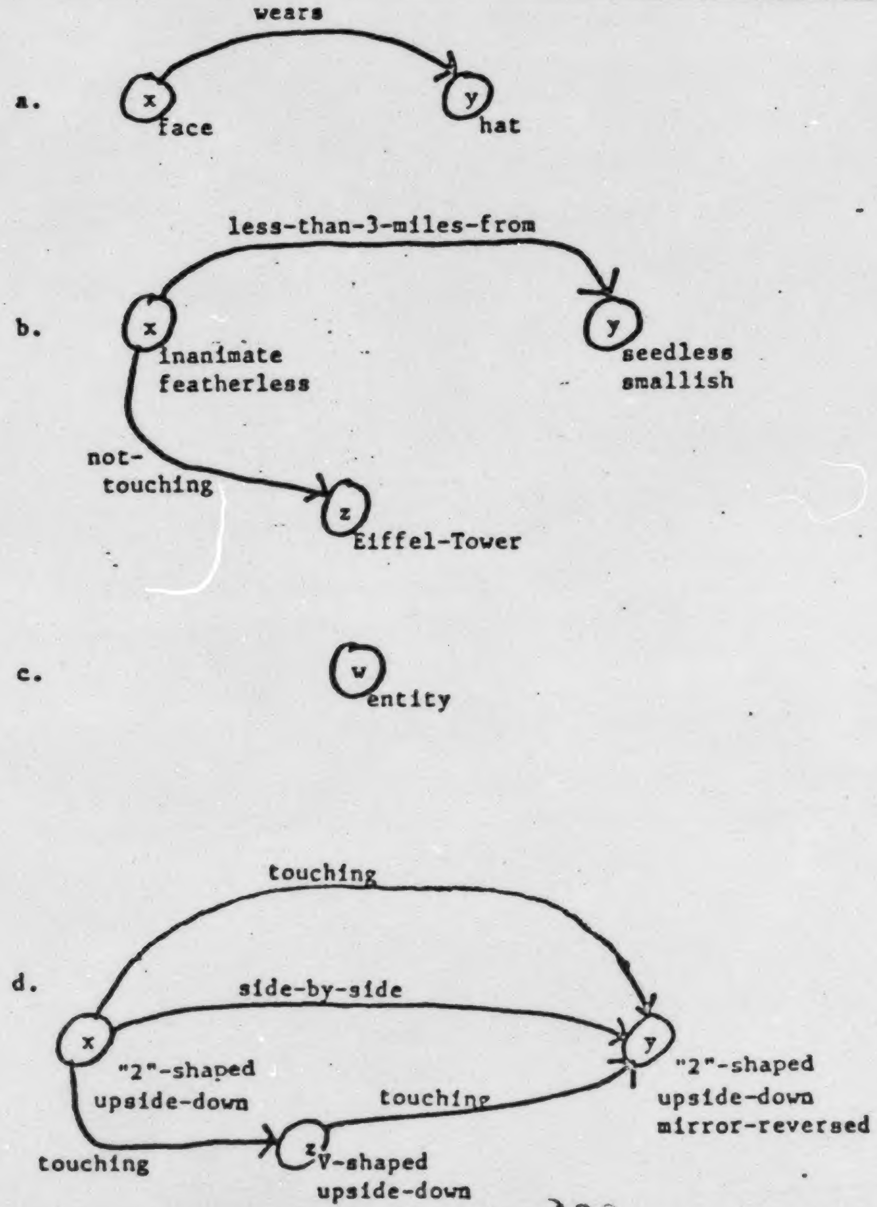

a.
wears
x
face
y
hat
b.
less-than-3-miles-from
x
inanimate
featherless
y
seedless
smallish
not-
touching
z
Eiffel-Tower
c.
w
entity
d.
touching
side-by-side
x
"2"-shaped
upside-down
y
"2"-shaped
upside-down
mirror-reversed
touching
touching
z
V-shaped
upside-down

Figure 4.

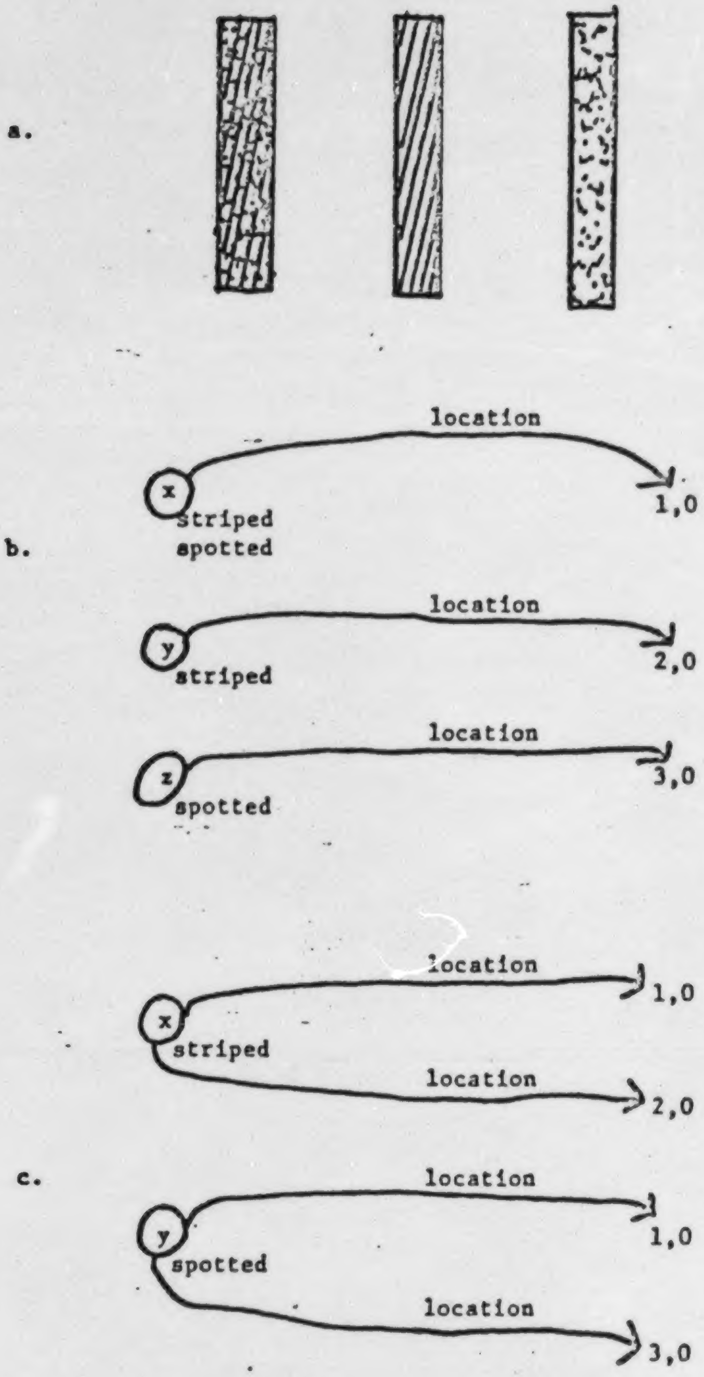

a.
b.
location
x
striped
spotted
1,0
location
y
striped
2,0
location
z
spotted
3,0
c.
location
x
striped
1,0
location
2,0
location
y
spotted
1,0
location
3,0

Figure 5.

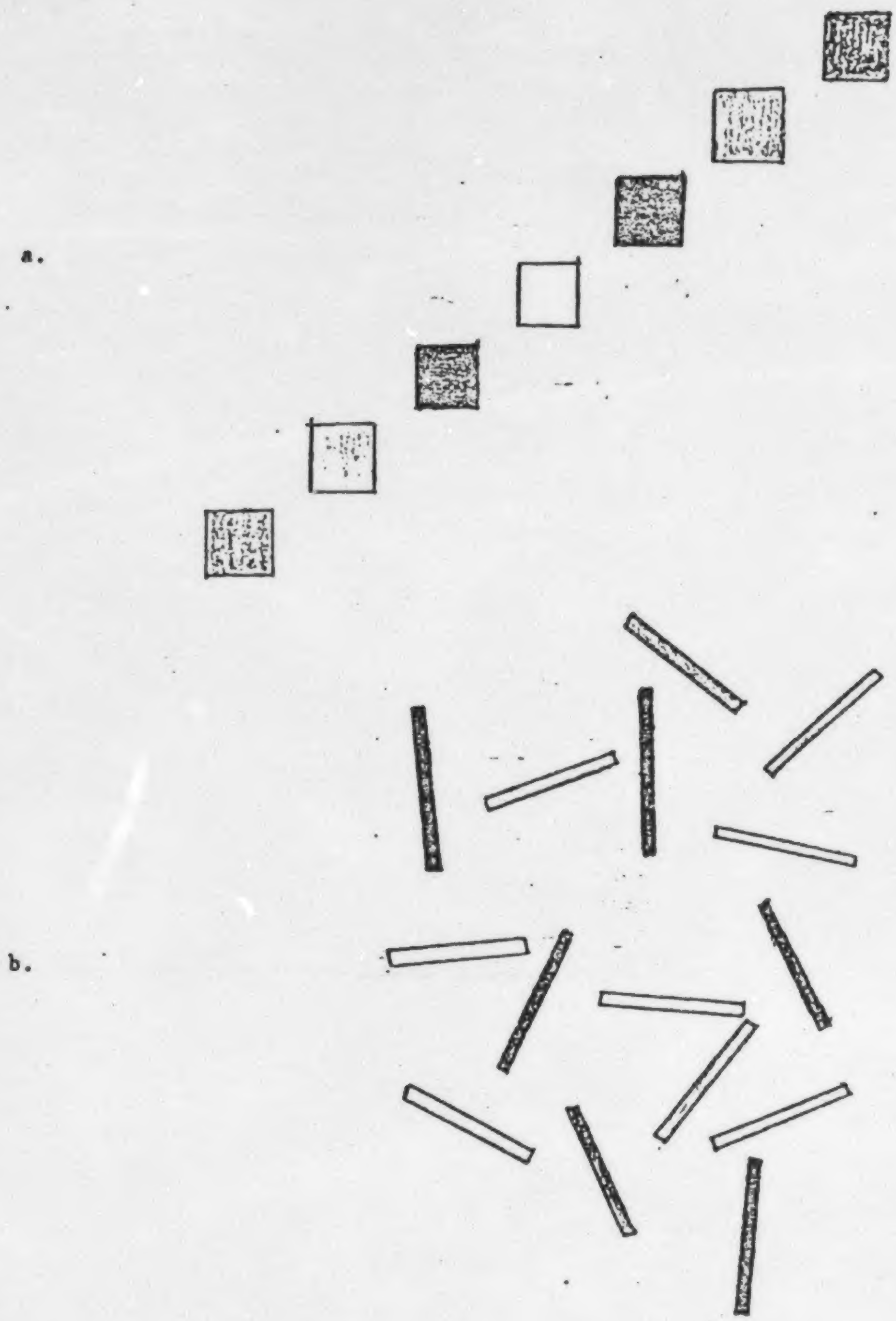


Figure 6.

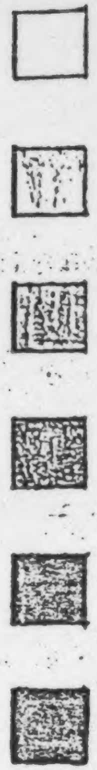


Figure 7.

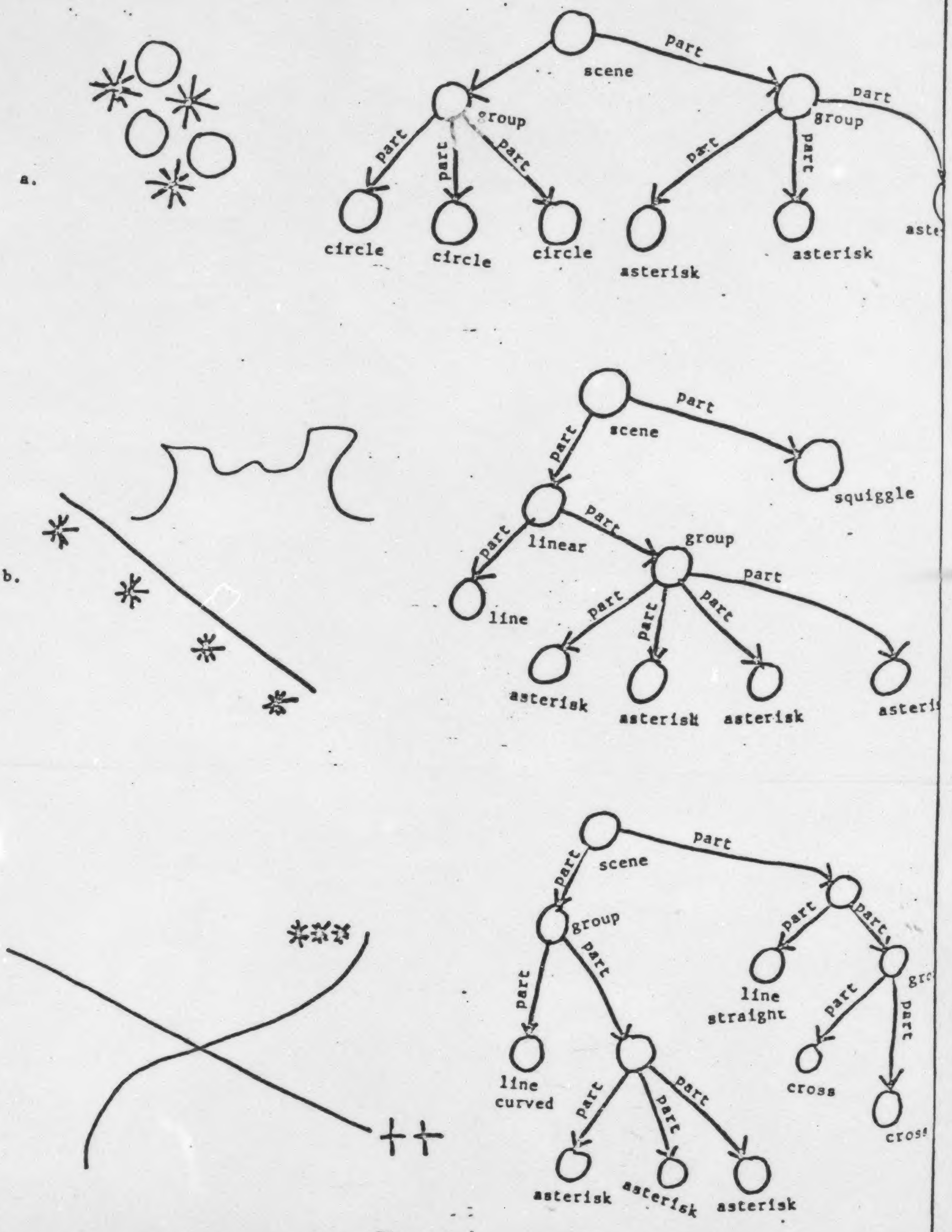

a.
scene
part
group
part
part
part
part
group
part
part
part
circle
circle
circle
asterisk
asterisk
aste
b.
scene
part
part
squiggle
linear
part
part
group
line
part
part
part
part
asterisk
asterisk
asterisk
asteri
c.
scene
part
part
group
part
part
part
part
line
straight
gro
part
part
line
curved
cross
cross
part
part
part
asterisk
asterisk
asterisk

Figure 8.

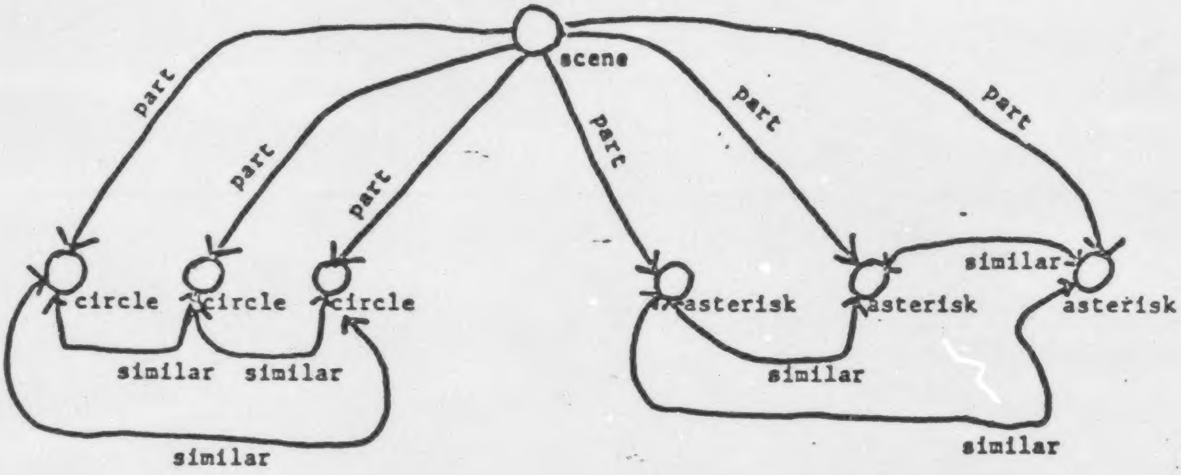

scene
part
part
part
part
part
part
circle
circle
circle
asterisk
asterisk
asterisk
similar
similar
similar
similar
similar
similar

Figure 9.

Figure 10.

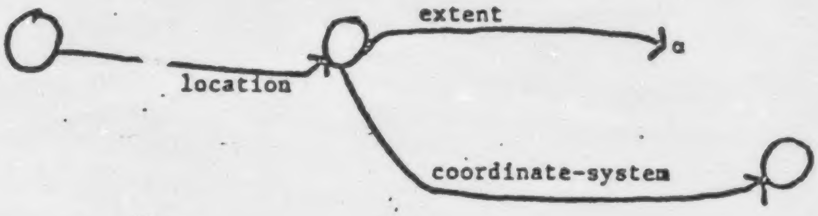
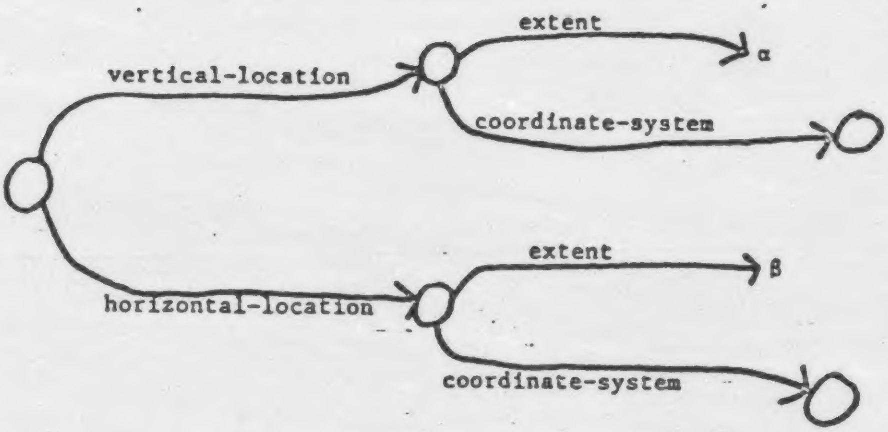

extent
α
location
coordinate-system

Figure 11.


extent
α
vertical-location
coordinate-system
extent
β
horizontal-location
coordinate-system

439

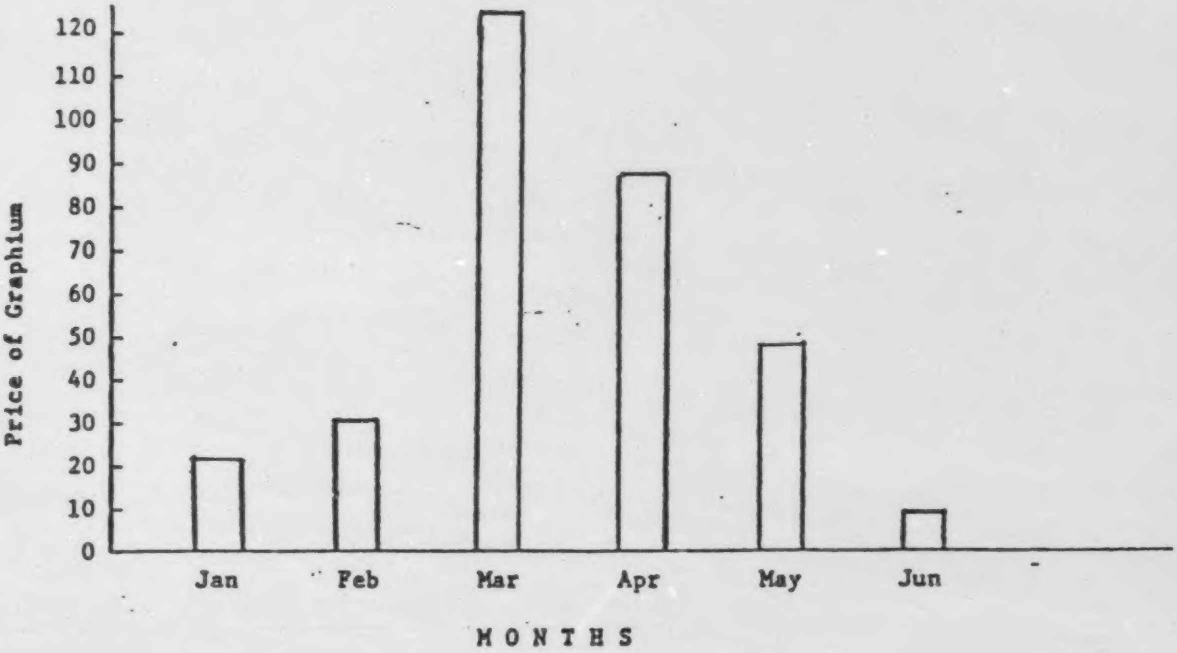

Price of Graphium
120
110
100
90
80
70
60
50
40
30
20
10
0
Jan
Feb
Mar
Apr
May
Jun
MONTHS

Figure 12.

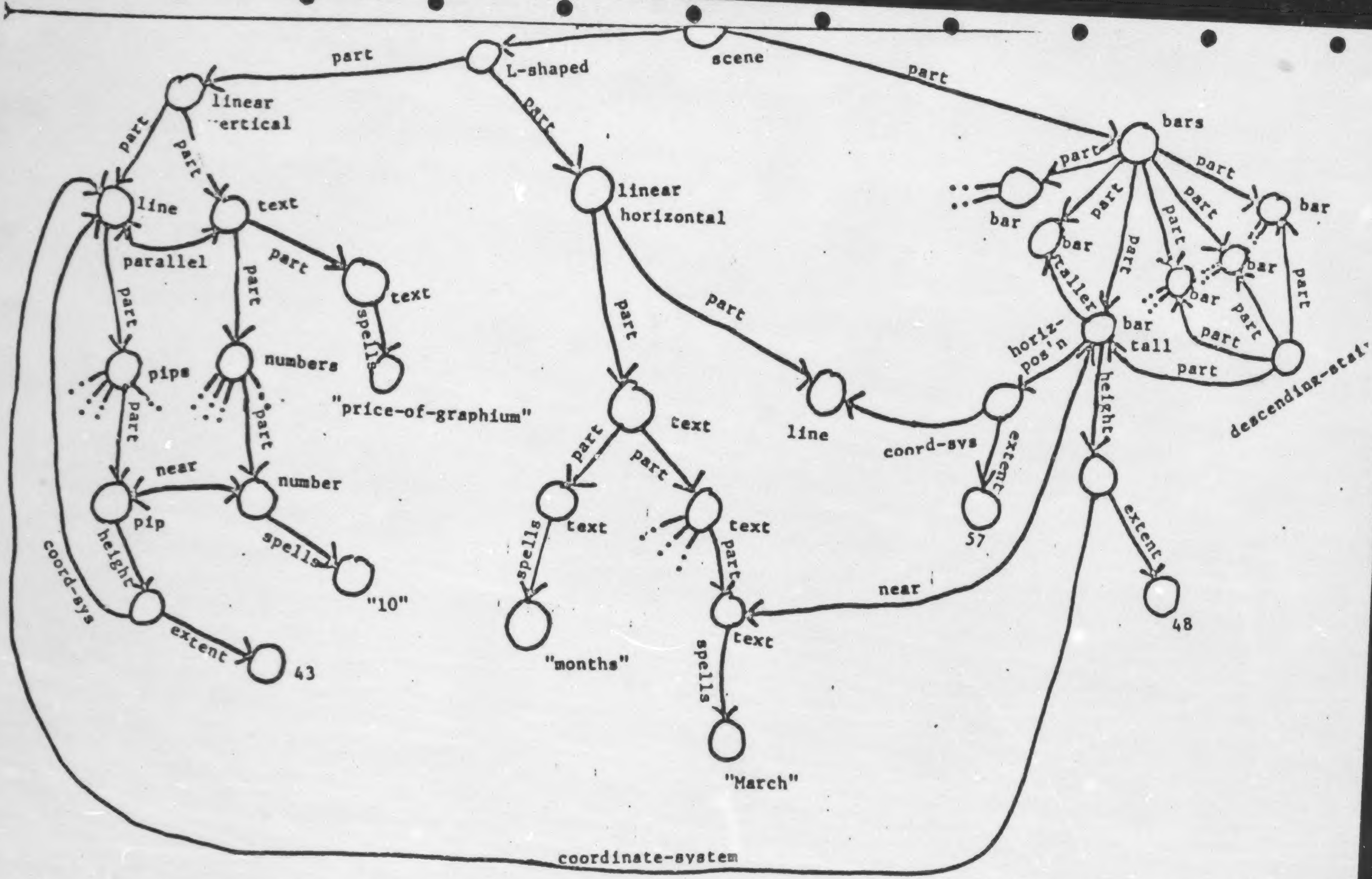

scene
part
L-shaped
part
bars
part
linear
vertical
part
part
linear
horizontal
line
text
parallel
part
part
part
text
spells
"price-of-graphium"
pips
part
numbers
part
pip
near
number
height
coord-sys
extent
43
spells
"10"
part
part
text
part
part
text
spells
"months"
text
part
text
spells
"March"
part
line
coord-sys
horiz-
pos'n
extent
57
part
bar
part
bar
part
part
part
bar
bar
bar
taller
bar
tall
part
part
part
part
descending-stai
height
extent
48
near
coordinate-system

Figure 13

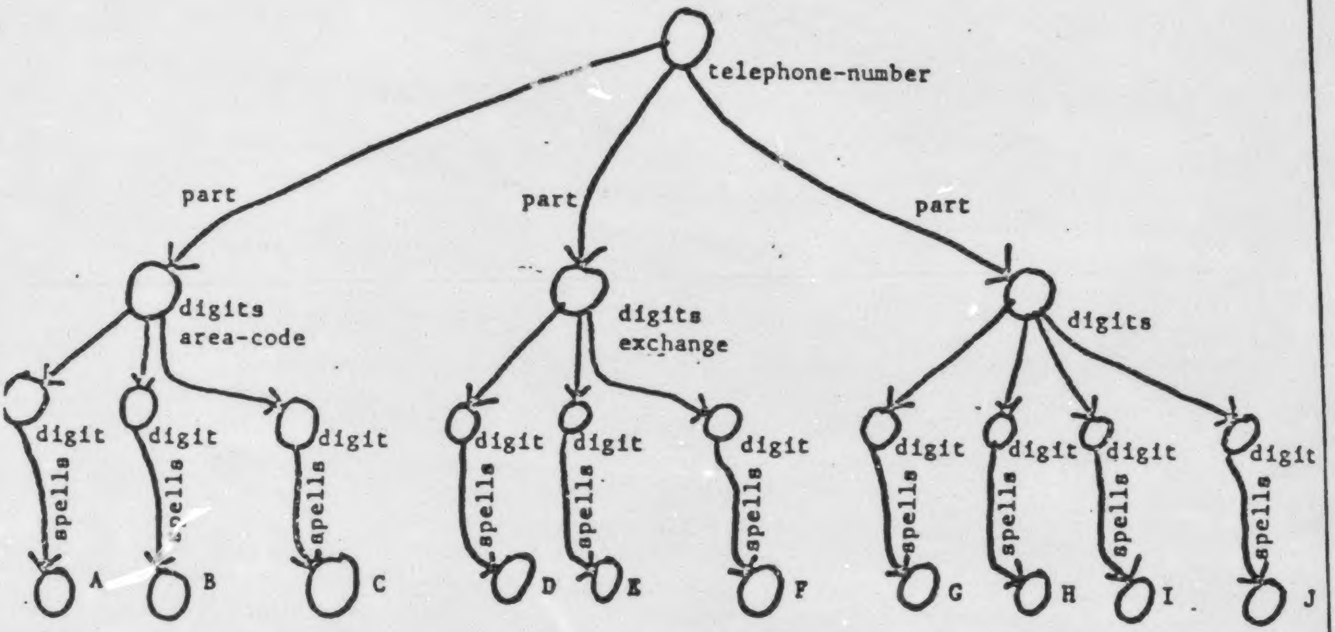

telephone-number
part
part
part
digits
area-code
digits
exchange
digits
digit
digit
digit
digit
digit
digit
digit
digit
digit
digit
spells
spells
spells
spells
spells
spells
spells
spells
spells
spells
A
B
C
D
E
F
G
H
I
J

Figure 15.

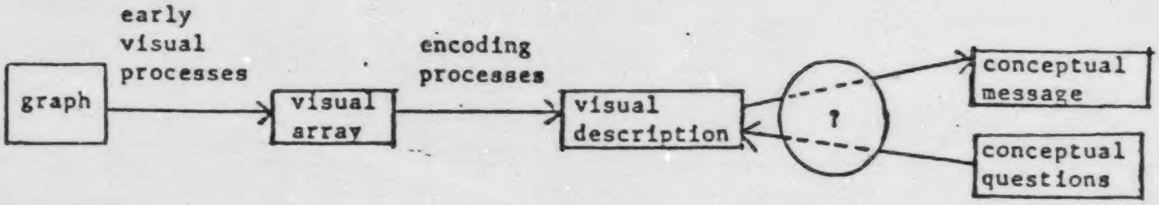

early
visual
processes
encoding
processes
graph
visual
array
visual
description
?
conceptual
message
conceptual
questions

Figure 14

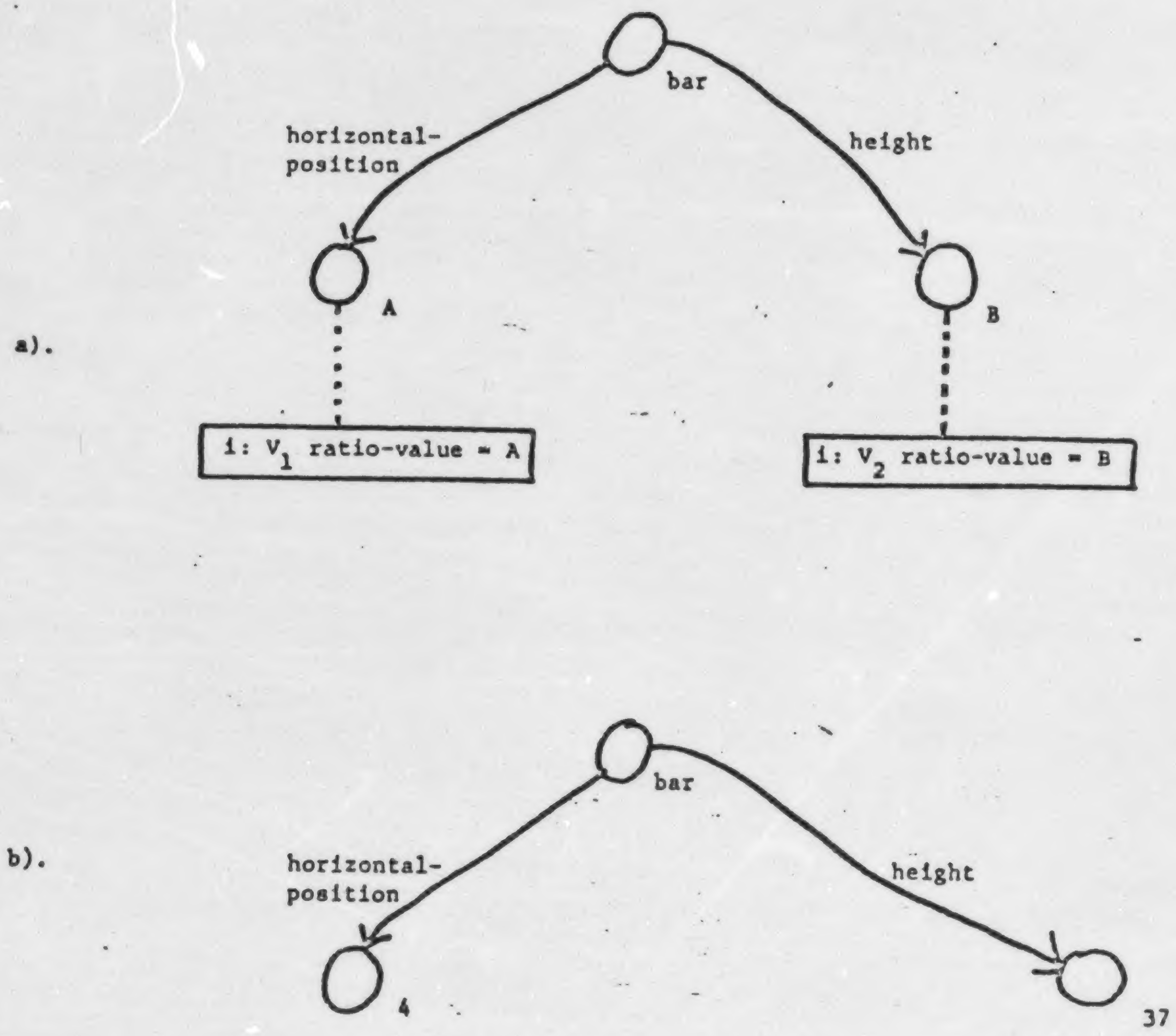

a).
bar
horizontal-position
height
A
B
i: V_1 ratio-value = A
i: V_2 ratio-value = B
b).
bar
horizontal-position
height
4
37

Figure 16.

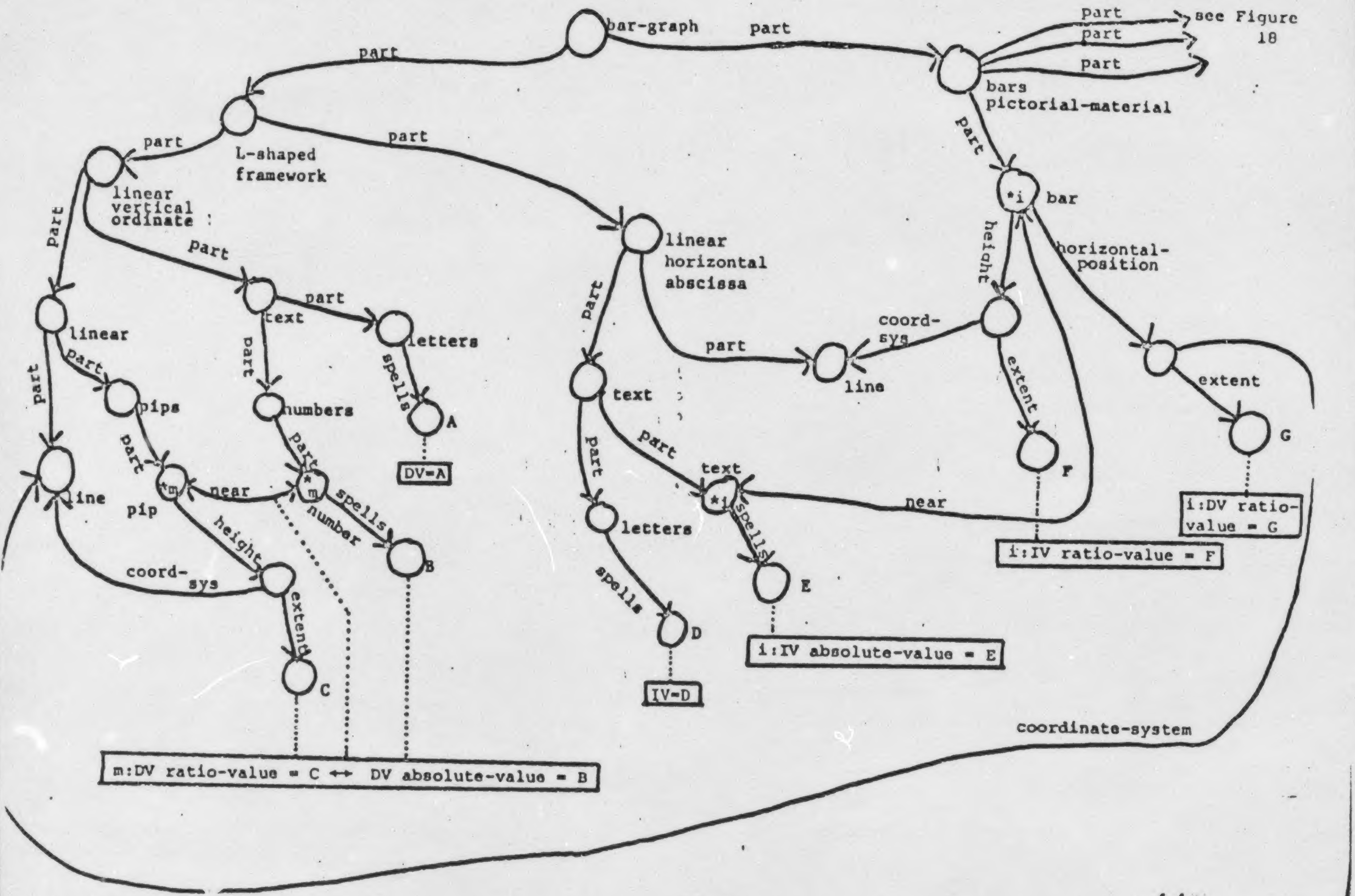

bar-graph
part
part
part
part
part
see Figure 18
bars
pictorial-material
L-shaped framework
part
part
linear vertical ordinate
part
part
linear
text
part
letters
spells
A
DV=A
part
part
pips
part
numbers
part
*m
pip
near
* m
number
spells
B
line
coord-sys
height
extent
C
m:DV ratio-value = C ↔ DV absolute-value = B
linear horizontal abscissa
part
text
part
letters
spells
D
IV=D
part
text
*i
spells
E
i:IV absolute-value = E
part
line
coord-sys
bar
*i
height
extent
F
i:IV ratio-value = F
horizontal-position
extent
G
i:DV ratio-value = G
near
coordinate-system

Figure 17.

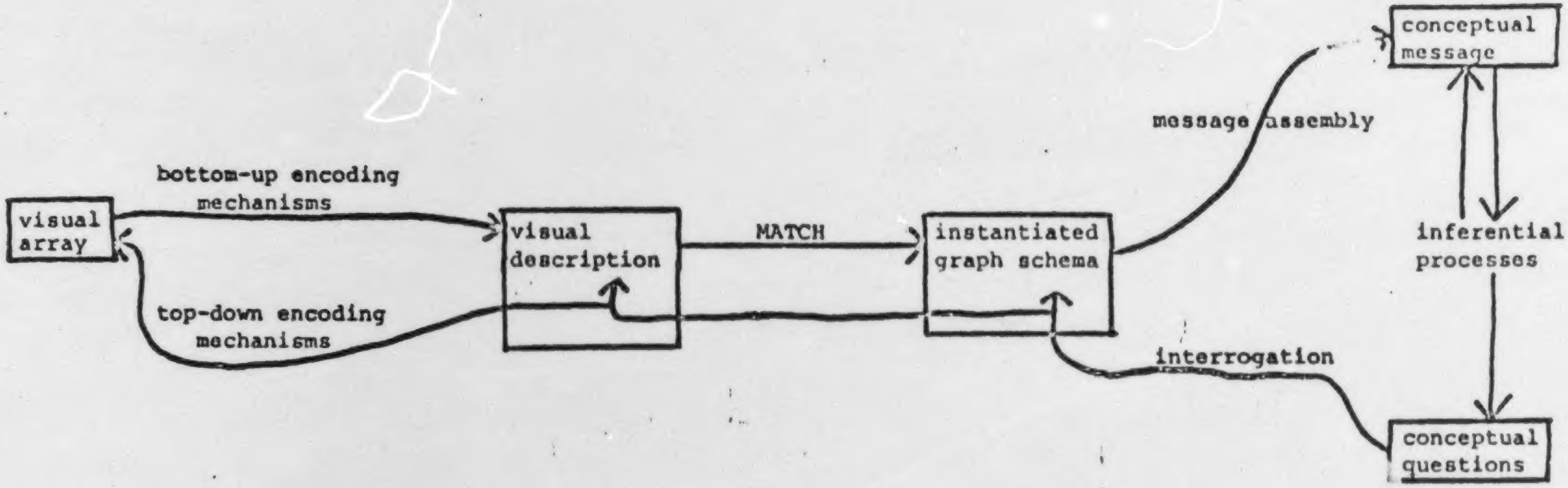

visual array
bottom-up encoding mechanisms
visual description
top-down encoding mechanisms
MATCH
instantiated graph schema
message assembly
conceptual message
inferential processes
interrogation
conceptual questions

Figure 19.

Figure 17.

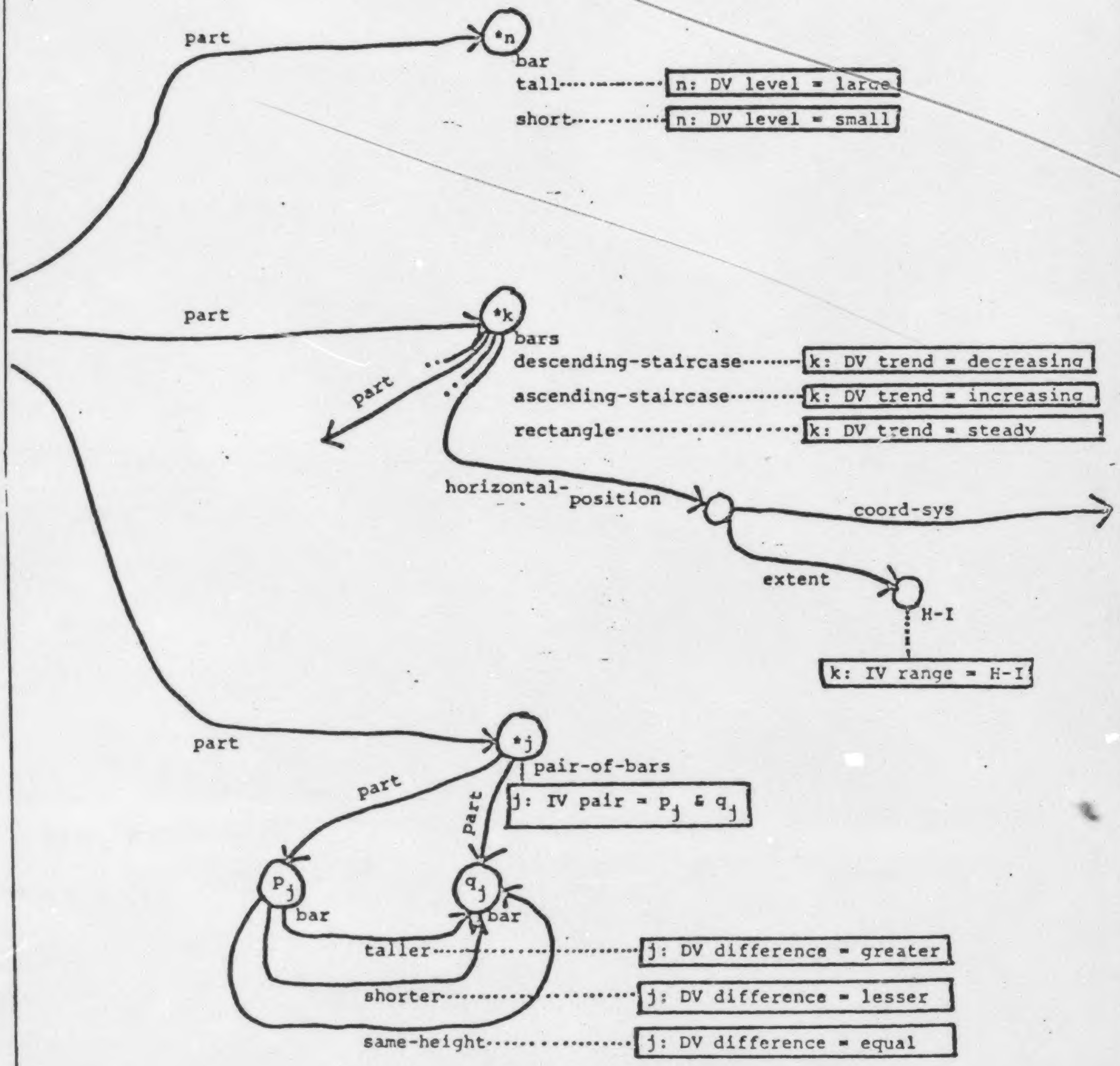

part
*n
bar
tall
n: DV level = large
short
n: DV level = small
part
*k
bars
descending-staircase
k: DV trend = decreasing
ascending-staircase
k: DV trend = increasing
rectangle
k: DV trend = steady
part
horizontal-position
coord-sys
extent
H-I
k: IV range = H-I
part
*j
pair-of-bars
j: IV pair = p_j & q_j
part
part
p_j
q_j
bar
bar
taller
j: DV difference = greater
shorter
j: DV difference = lesser
same-height
j: DV difference = equal

Figure 18.

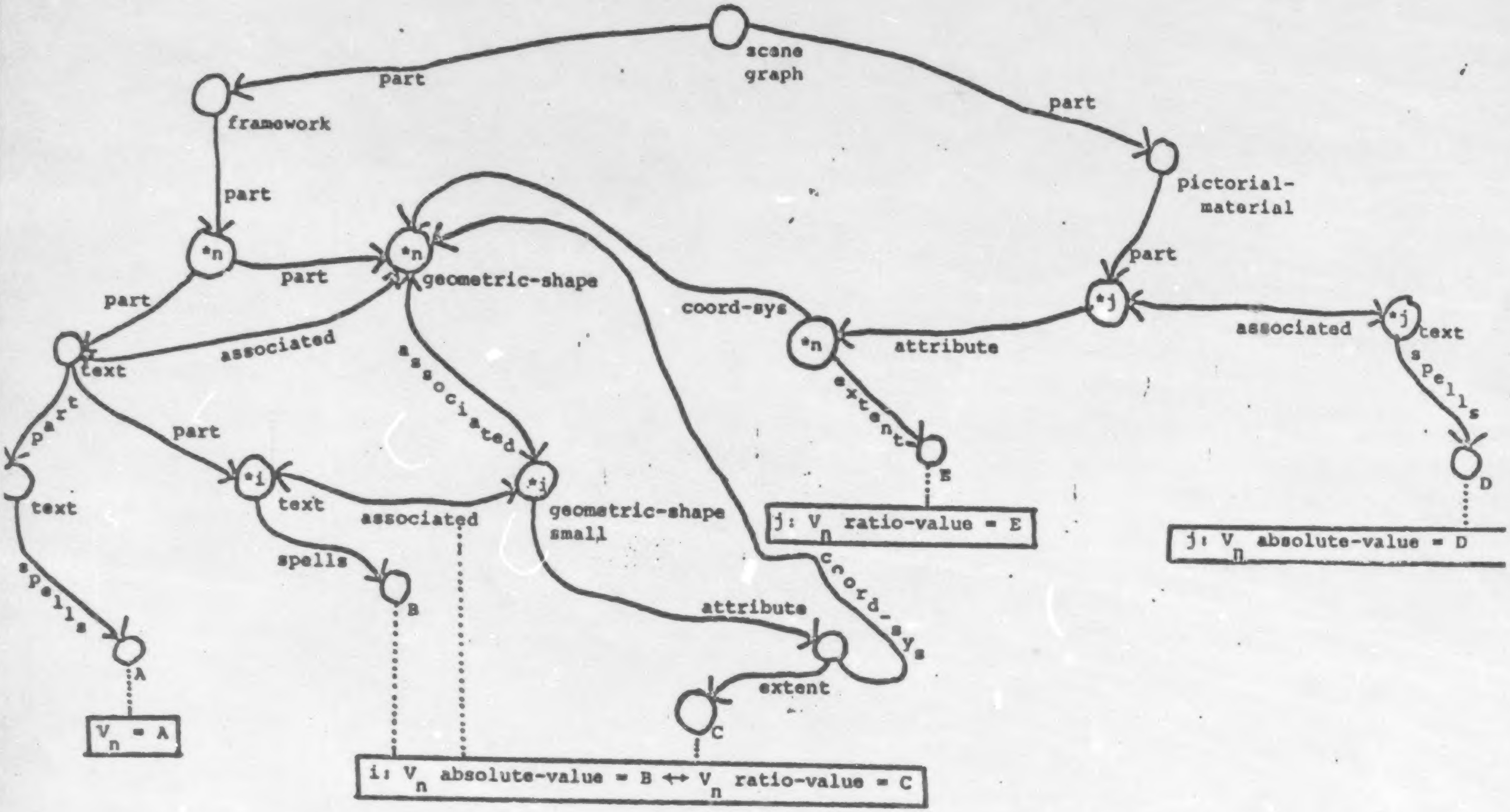

scene
graph
part
framework
part
part
pictorial-
material
part
*n
part
*n
geometric-shape
part
coord-sys
*j
associated
*j
text
*n
attribute
text
associated
associated
spells
extent
part
part
E
D
text
*i
text
associated
*j
geometric-shape
small
j: V_n ratio-value = E
j: V_n absolute-value = D
spells
spells
B
coord_sys
attribute
A
extent
C
V_n = A
i: V_n absolute-value = B ↔ V_n ratio-value = C

Figure 20.0

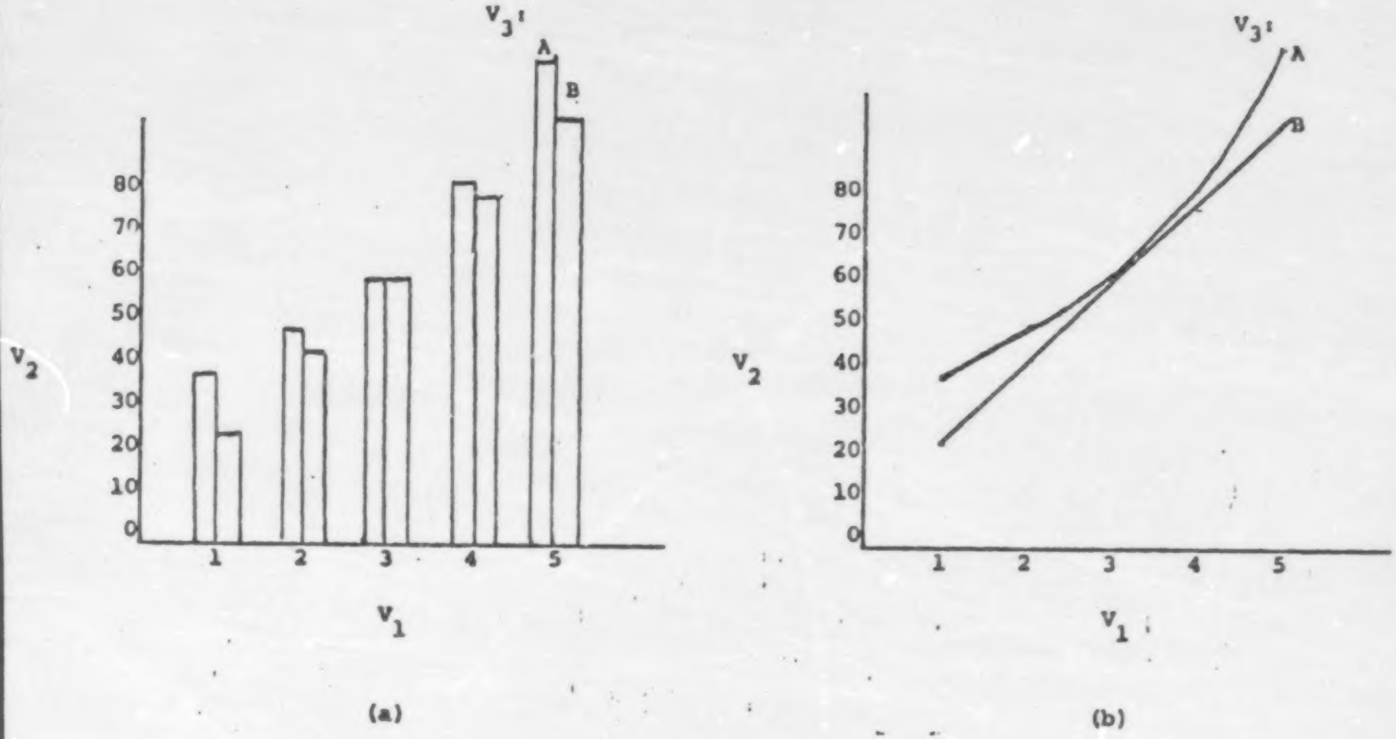
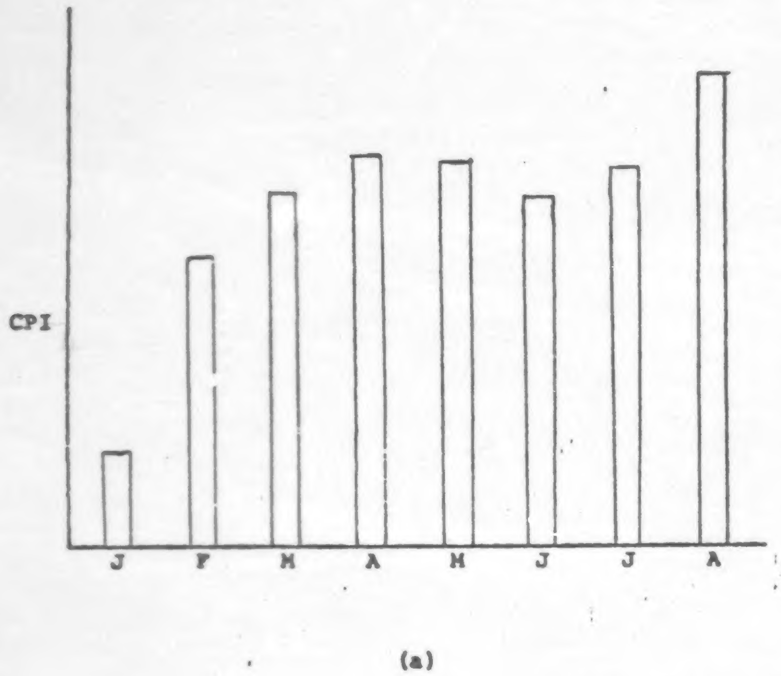


Figure 21.

453



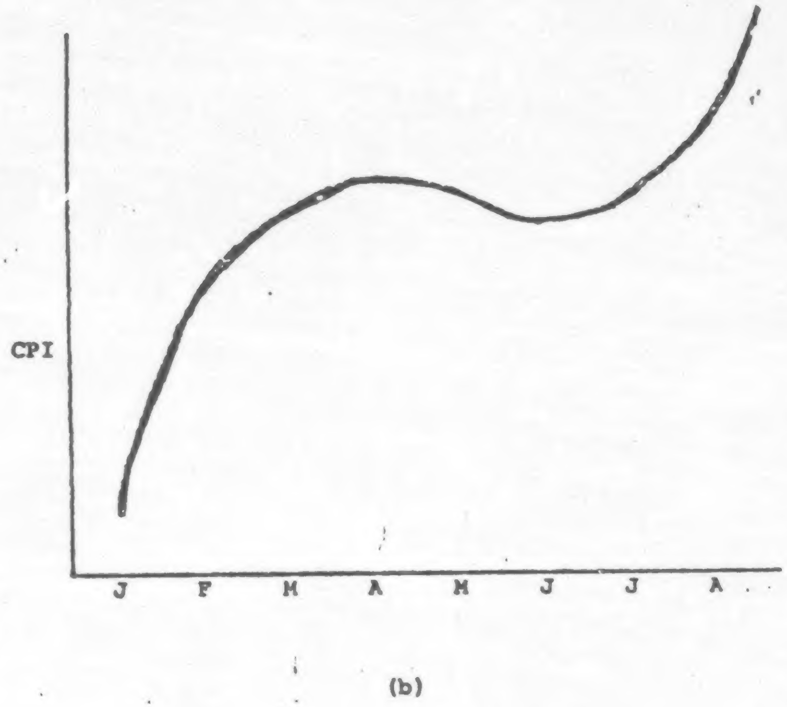


Figure 22.

454

455

set_1 = A

set_2 = B

set_3 = C

set_1 intersects set_2

set_2 superset-of set_3

set_1 disjoint-from set_3

Figure 23(d).

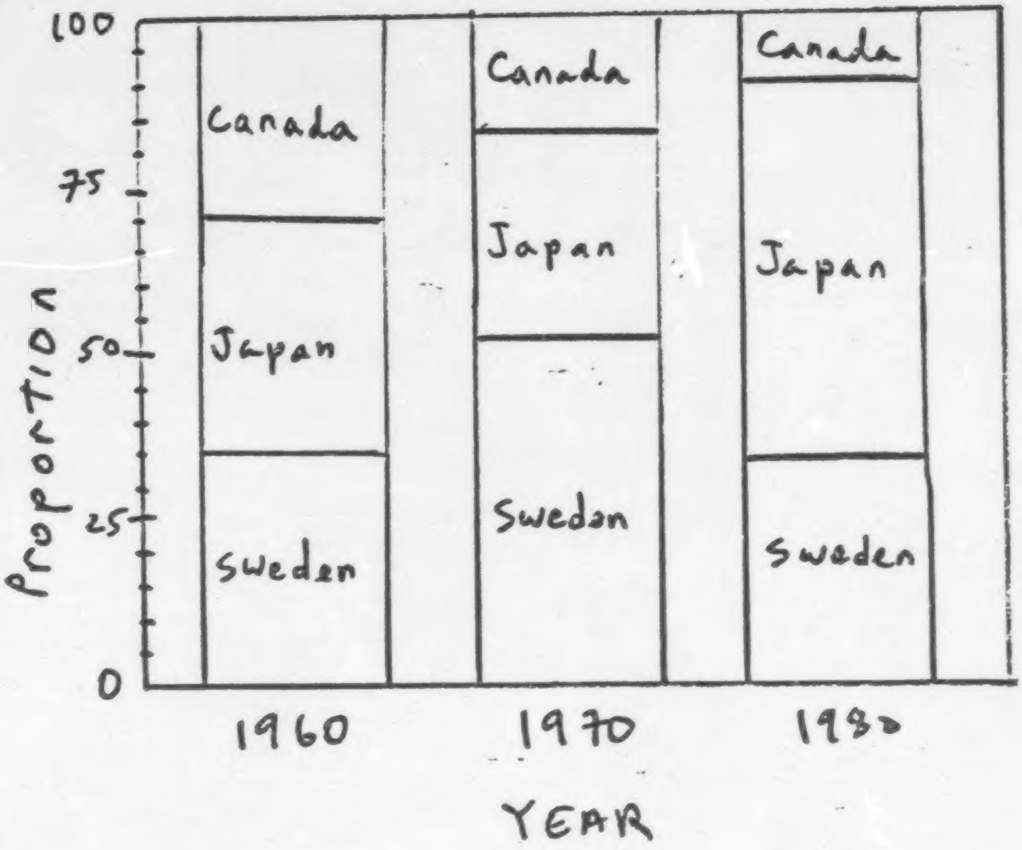


Fig 7.1 Proportion of mental illness in population
(To pg 16)

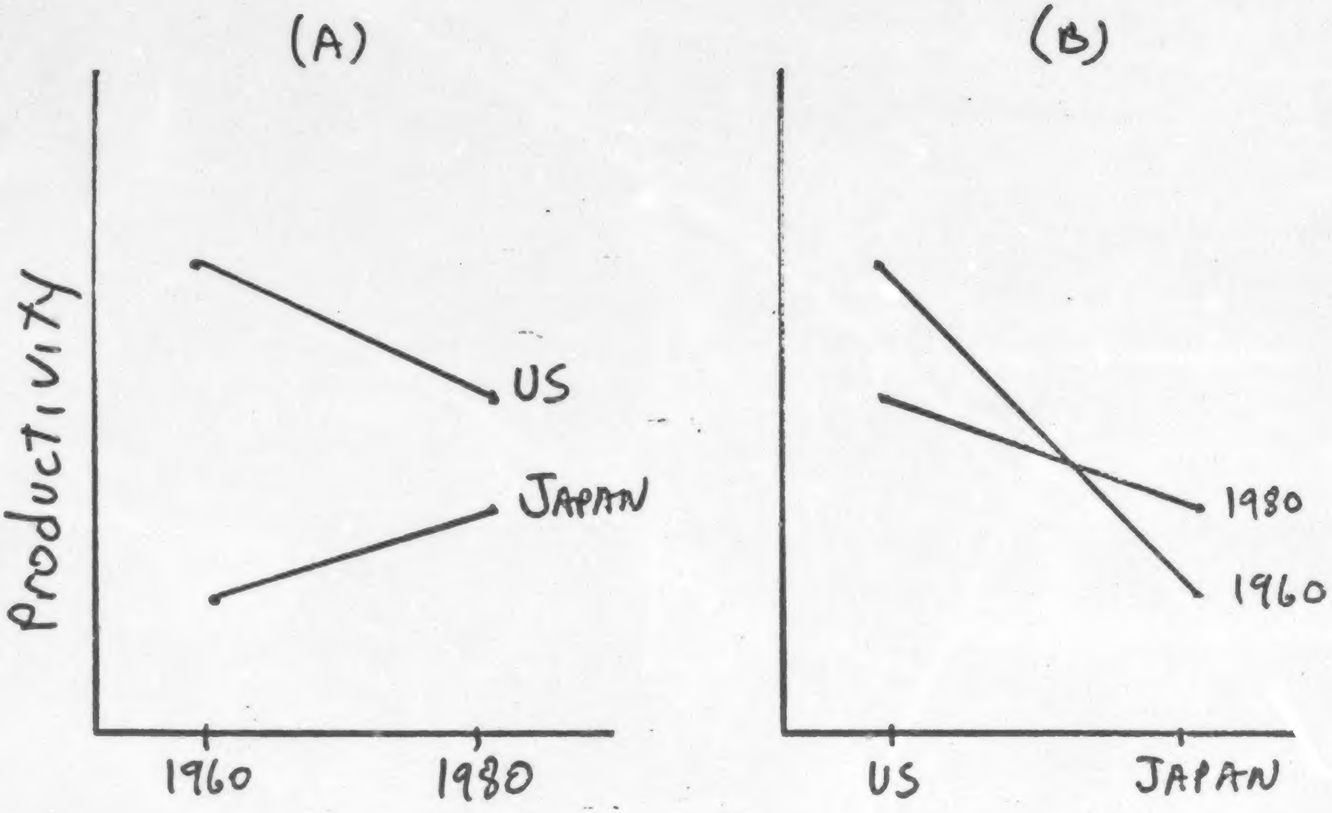


Fig 7.2. Two ways of graphing the same data, changing foreground and background.

(p 18)

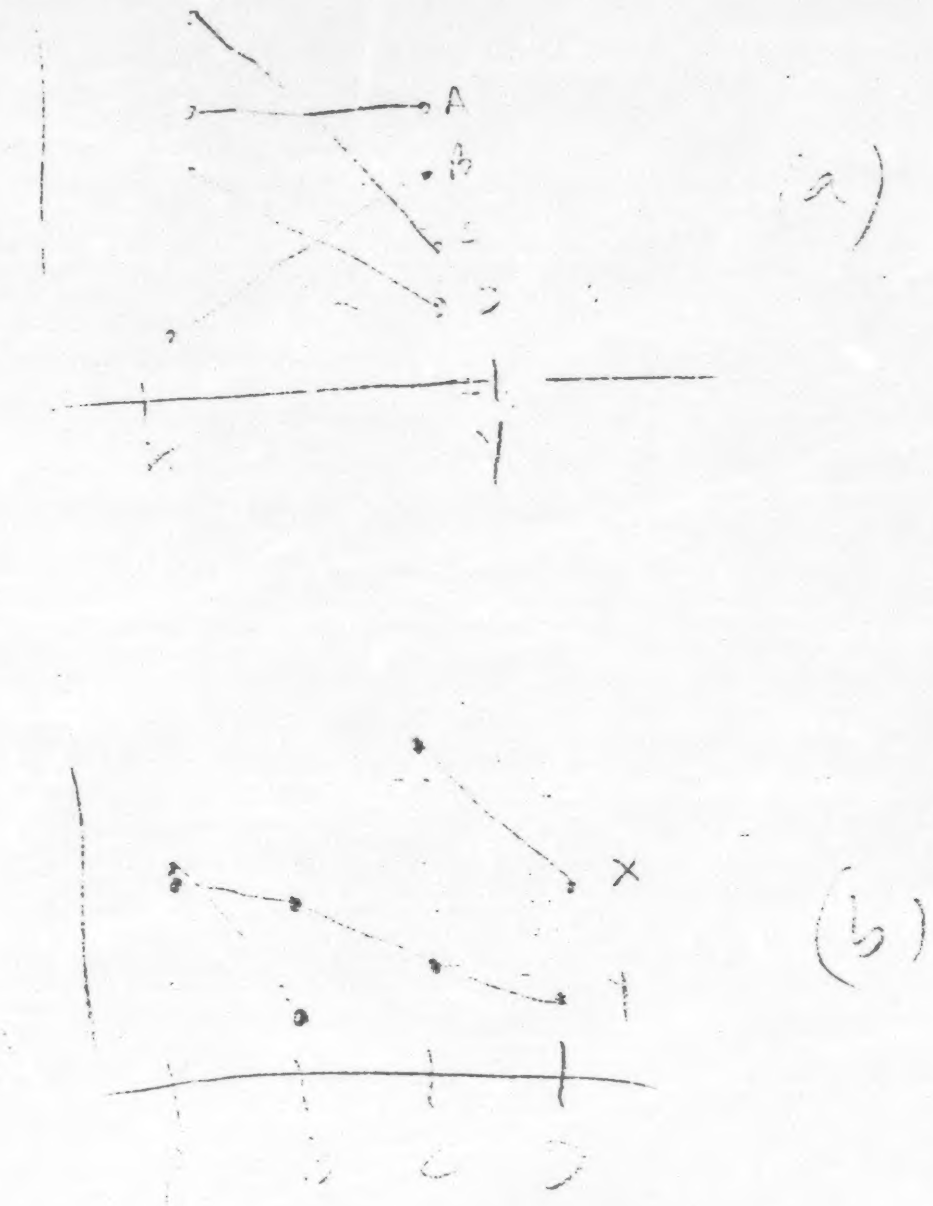


Fig 7.3

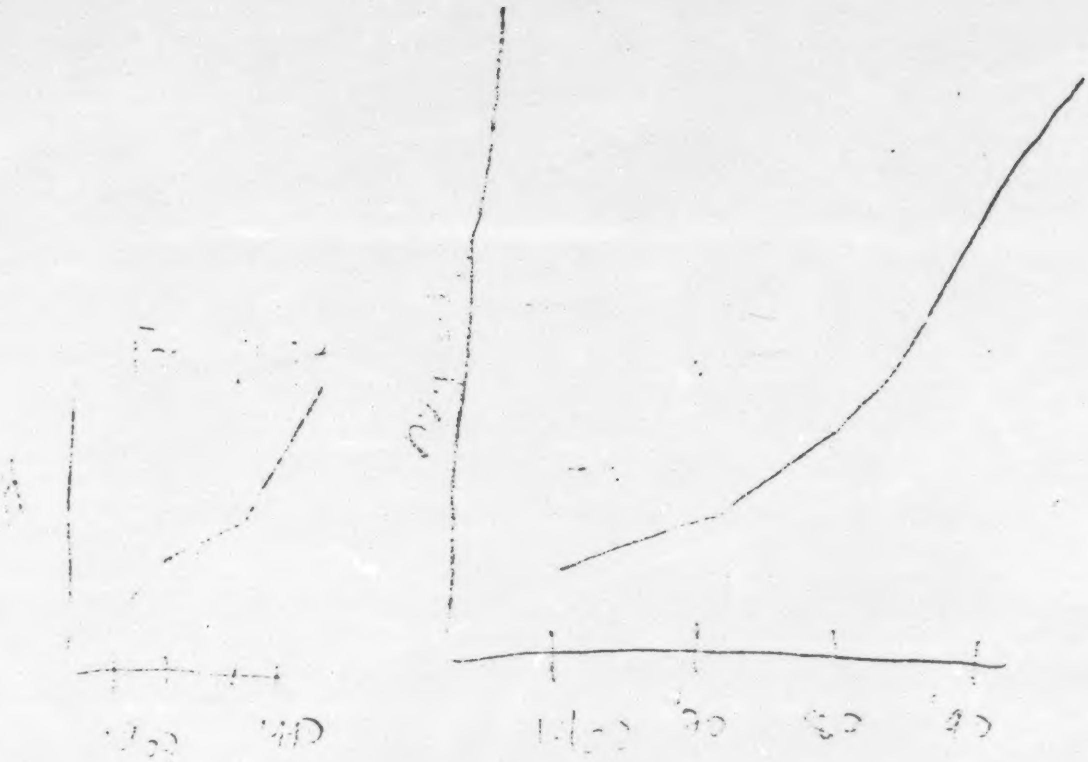


Figure 7.4

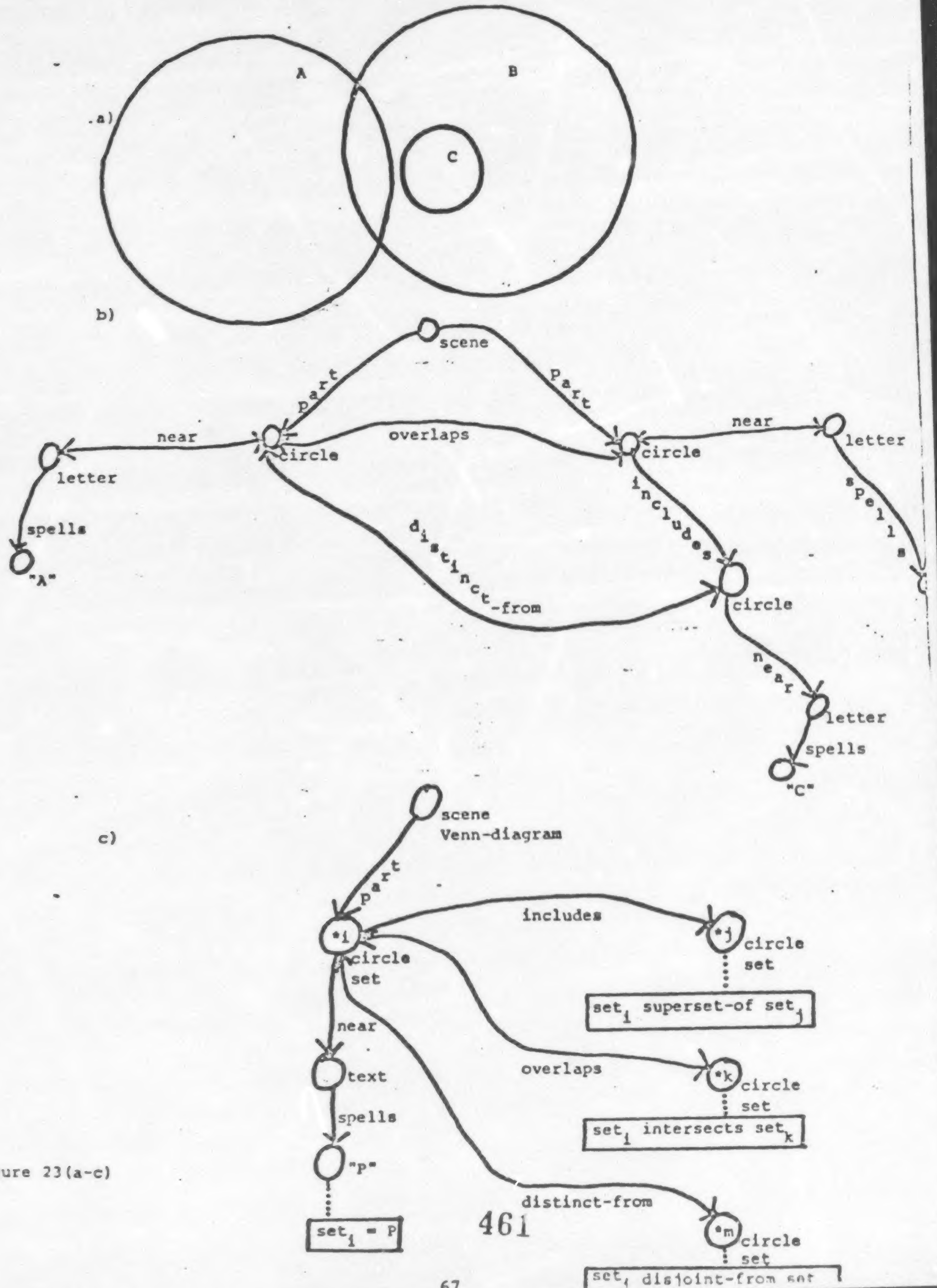

a)
A
B
C
b)
scene
part
part
near
overlaps
near
letter
circle
circle
letter
spells
includes
spells
"A"
distinct-from
circle
near
letter
spells
"C"
c)
scene
Venn-diagram
part
includes
*i
circle
set
*j
circle
set
set$_i$ superset-of set$_j$
near
text
overlaps
*k
circle
set
spells
set$_i$ intersects set$_k$
"P"
distinct-from
set$_i$ = P
*m
circle
set
set$_i$ disjoint-from set

Figure 23(a-c)

Table ~~1~~ 8.1. Species of visual displays. Examples are in parentheses.

	TYPE		
USE	Intrinsic Configuration	Model	Symbol
Illustration	(diagram of rhombus)	(floorplan)	(graphs)
Problem Solving	(diagram used to prove theorem)	(pulley diagram used to anticipate movements)	(Venn diagram)
Problem Defining	(Euclidean experimental sketch)	(Einsteinian "thought experiment" image)	(sketch of Hilbert space)

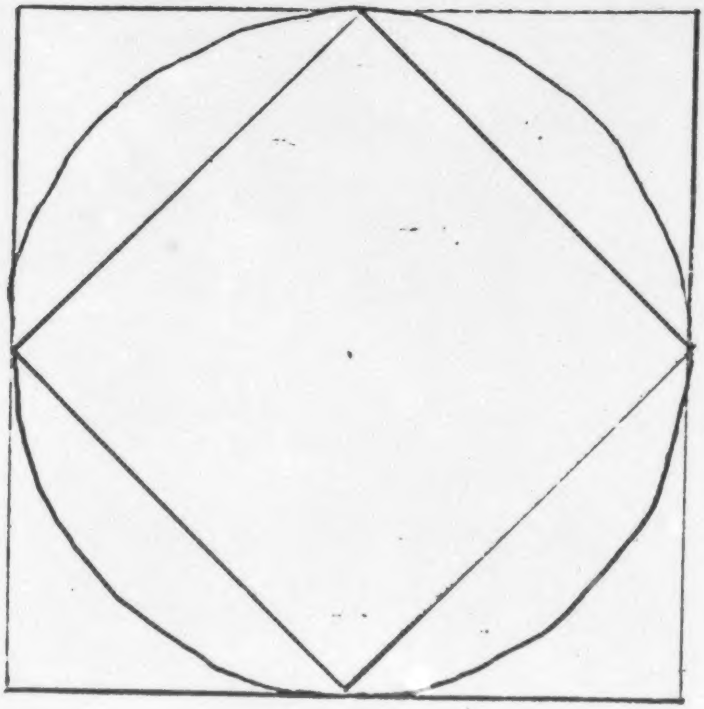


fig 8.1

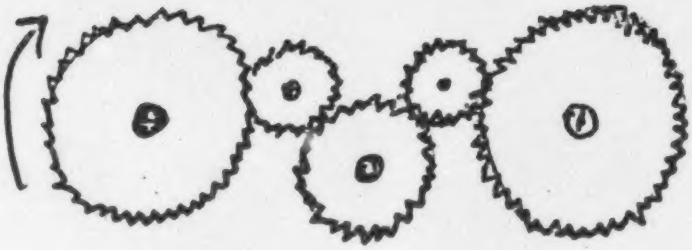


Fig 8.2